# A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition

Daniel Weinland, Rémi Ronfard, Edmond Boyer

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition*

Daniel Weinland — Remi Ronfard — Edmond Boyer

**N° 7212**

Février 2010

Vision, Perception and Multimedia Understanding


ISSN 0249-6399 ISRN INRIA/RR--7212--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

# A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition

Daniel Weinland*, Remi Ronfard†, Edmond Boyer‡

Theme : Vision, Perception and Multimedia Understanding
Perception, Cognition, Interaction
Équipes-Projets Lear et Perception

Rapport de recherche n° 7212 — Février 2010 — 54 pages

**Abstract:** Action recognition has become a very important topic in computer vision, with many fundamental applications, in robotics, video surveillance, human computer interaction, and multimedia retrieval among others. The number of works published is steadily increasing, and action recognition is meanwhile presented with numerous publications at recent conferences. A large variety of approaches have been described. The purpose of this survey is to give an overview and categorization of the approaches used. We concentrate on approaches that aim on classification of full-body motions, such as kicking, punching, waving, etc. and we categorize them according to how they represent the spatial and temporal structure of actions; how they segment actions from an input stream of visual data; and how they learn a view-invariant representation of actions.

**Key-words:** computer vision, action recognition

* Deutsche Telekom Laboratories, TU Berlin, Germany
† INRIA Team Lear, Grenoble, France
‡ INRIA Team Perception, Grenoble, France

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

# Etat de l'Art sur les Méthodes basées Vision en Représentation, Segmentation et Reconnaissance d'Actions

**Résumé :** La reconnaissance d'actions est un problème important en vision par ordinateur, avec de nombreuses applications fondamentales en robotique, télé-surveillance, interaction homme-machine et indexation multimedia, entre autres. Le nombre de publications sur ce sujet augmente régulièrement dans les conférences du domaine. Une grande variété d'approches ont été décrites. Le but de ce rapport est de dresser un état de l'art du domaine et de proposer une classification des approches utilisées. Nous nous focalisons sur le problème de la classification des actions faisant intervenir l'ensemble du corps, telles que s'asseoir, se lever, battre des mains, donner un coup de pied ou un coup de poing, etc. Nous classons les différentes approches du problème en fonction des représentations spatiales et temporelles qu'elles donnent des actions; de la façon dont elles permettent de segmenter les actions dans un flux visuel continu; et de leur capacité à apprendre des modèles indépendants du point de vue.

**Mots-clés :** reconnaissance d'actions, vision par ordinateur

# Contents

# 1 Introduction

Action recognition is a very active research topic in computer vision with many important applications, including human-computer interfaces, content-based video indexing, full-video search, video surveillance, robotics, programming by demonstration, among others. Historically, visual action recognition has been divided into sub-topics such as gesture recognition (especially hand gestures) for human-computer interfaces [36, 122], facial expression recognition [204], and movement behavior recognition for video surveillance [66]. However full-body actions usually include different motions and require a unified approach for recognition, encompassing facial actions, hand actions and feet actions.

Action recognition is the process of naming actions, usually in the simple form of an action verb, using sensory observations. Technically, an action is a sequence of movements generated by a human agent during the performance of a task. As such, it is a four-dimensional object, which may be further decomposed into spatial and temporal *parts*. In this paper, we are only concerned with visual observations, typically by means of one or more video cameras, but it should be noted that actions can of course also be recognized from other sensory channels, including audio. An action label is a name, such that an average human agent can understand and perform the named action. The task of action recognition is to name actions, i.e. determine the action label that best describes an action instance, even when performed by different agents under different viewpoints, and in spite of large differences in manner and speed. A typical set-up for testing and evaluating action recognition systems consist in sending instructions to the actors, using simple action verb imperatives, and to compare them with the recognized action names.

To reach that goal, the various approaches typically employ a combination of *vision* and *machine learning* tools. Vision techniques attempt to extract action discriminative features from the video sequences, while providing appropriate robustness to distracting cues. Machine learning attempts to learn statistical models form those features, and to classify new features based on the learned models. Two issues which are thereby of particular importance are to deal with changing viewpoints and to segment the observed motions into semantic meaningful instances of actions.

Note that our definition of an action is more restrictive than the one proposed by Pinhanez [127] when he states that actions are sequences of movements performed in a given context ($action = context + movement$), with the example of *typing* or *playing piano* which involve the same quick movements of the fingers in the different contexts of a computer desk or a concert hall. For our purpose, they are one and the same *action* of quickly moving one's fingers, and this action can be executed as part of different *tasks*, such as playing the piano or typing. The importance of context for visual action recognition is the focus of an excellent recent survey on the *meaning of action* [85]. Here, we concentrate on the *structure of action* by reviewing vision-based techniques that can be used for analyzing, segmenting and classifying movements in order to recognize actions independently of the task and context where it is performed.

Generic action recognition has already been surveyed in [25, 2, 49, 106, 107] in the context of motion capture and body tracking, and in [66] in the context of surveillance. High-level analysis of activities was recently surveyed in [174]. In contrary, our survey focuses exclusively on action recognition, and it is the

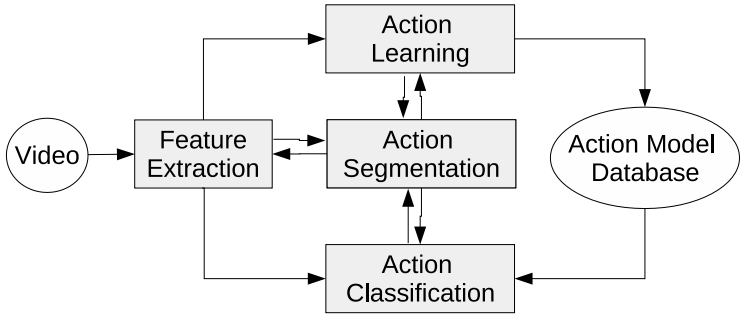



Figure 1: A typical data-flow for generic action recognition system comprises inter-dependent stages of feature extraction, learning, segmentation and classification.

first work investigating the three related issues of representing, segmenting and recognizing actions.

Figure 1 illustrates the major components of a generic action recognition system and their typical arrangement.

**Feature extraction** is the main vision task in action recognition and consist in extracting posture and motion cues from the video that are discriminative with respect to human actions. Very different representations can be used, ranging from complex body models to simple silhouette images. In either case, issues such as person location, robustness to partial occlusion, background clutter, shadows and different illumination need to be addressed. Further representations should provide some insensitivity to different types of clothing and physiques.

**Action learning and classification** are the steps of learning statistical models from the extracted features, and using those models to classify new feature observations. A major challenge thereby is to deal with the large variability that an action class can exhibit, in particular if performed by different subjects of different gender and size, and with different speed and style. Action categories which might seem clearly defined to us, such as kicking, punching, or waving, for instance, can have very large variability when performed in practice. It is thus a particular challenge to design an action model, which identifies for each action the characteristic attitudes, while maintaining appropriate adaptability to all forms of variations.

**Action segmentation** is necessary to cut streams of motions into single action instances that are consistent to the set of initial training sequences used to learn the models. Closely related are the questions: how to choose such initial segmentations; and is there something like an elementary vocabulary of primitive motions in action articulation and perception?

Vision-based techniques for representing, segmenting and recognizing human actions can be classified according to many different criteria, e.g. the body parts involved (facial expressions, hand gestures, upper-body gestures, full-body gestures, etc.); the selected image features (interest points, landmarks, edges, optical flow, etc.); the class of statistical models used for learning and recognition (nearest neighbors, discriminant analysis, Markov models, Bayesian networks, conditional random fields, etc.). The classification we have found to be the

most useful is how the different methods proposed in the literature represent the spatial and temporal structure of actions. Indeed, our analysis of the recent literature in computer vision reveals a large variety of approaches in both the temporal and the spatial dimensions, which can be summarized as follows. In the spatial domain, action recognition can be based on global image features, aligned to the geometry of the scene or camera; or on parametric image features, aligned to the geometry of the human body; or on local image features, without structure. We review those three important classes in Section 2. In the temporal domain, action recognition can be based on global temporal signatures, such as stacked features, that represent an entire action from start to finish; or on grammatical models that represent how the moments of actions are organized sequentially, usually with several states and transitions between those states; or on sparse and unstructured observations, such as isolated key-frames. We review those three important classes in Section 3. By combining the three main spatial classes with the three main temporal classes, we end up with a synoptic classification of action recognition into nine basic classes, shown in Table 1.

Additional difficulties are introduced when we allow to observe actions from different and changing views. In such unconstrained realistic settings a single pose or motion can result in an almost infinite number of possible observations. An appropriate representation needs thus to account for such changes. To this aim, view-independent approaches have been introduced. Because of the importance of that issue and because of the large variety of different approaches that have been proposed, we discuss those approaches in a separate section.

The paper is therefore organized as follows. First, we present a general overview of action recognition methods, based on how they represent the spatial structure of actions in Section 2, and the temporal structure of actions in Section 3. Then, we review the special topics of action segmentation in Section 4 and view-invariant action recognition in Section 5. We close this survey with a discussion on available datasets and experimental evaluation.

## 2 Spatial Action Representations

We begin this survey with a review of spatial representation used to discriminate actions from visual data. As mentioned previously, a first step in action recognition is the extraction of image features that are discriminative with respect to posture and motion of the human body. Various representations have been suggested. They mainly contrast by the amount of high level information they represent versus how efficient they are to extract in practice. For the purpose of this survey, we classify them into three main groups - body models, image models, and unstructured features. *Body models* are based on a parametric representation of the human body recovered from images using body-part detection and tracking. *Image models*, are based on dense image features computed over a regular grid. *Sparse features* are based on sparse image features computed at specially detected interest regions and loosely organized into a *spatial bag-of-features*.

Table 1: Classification of Action Recognition Methods based on Spatial (vertical axis) and Temporal Representations (horizontal axis). Only some of the more recent approaches are listed in each cell.

| | Parametric,<br>Action Grammar | Global,<br>Action Template | Local,<br>Bag of Features |
|---|---|---|---|
| Parametric, Body Model | Body Grammar<br>e.g.<br><br>Wang[183], Kojima[84], Zhao[203], Park[121], Ramanan[133], Green[54], Nguyen[111], Guerra-Filho[57], Parameswaran[118], Peursum[125], Kitani[82], Lv[99], Wang[184], Ali[4], Ikizler[67], Morency[108] | Body Template<br>e.g.<br><br>Guo[58], Niyogi[114], Gavrila[47], Seitz[153], Yacoob[195], Ben-Arie[8], Rao[136], Gritai[55], Alon[5], Sheikh[156], Yilmaz[200], Shen[157] | Bag of Postures<br>e.g.<br><br>— |
| Global, Image Model | Image Grammar<br>e.g.<br><br>Brand[17], Elgammal[35], Cuzzolin[28], Ogale[116], Robertson[138], Sminchisescu[161], Ahmad[3], Lv[100], Turaga[176], Weinland[187], Natarajan[110], Vitaladevuni[180] | Image Template<br>e.g.<br><br>Pierobon[126], Roh[141], Weinland[190], Kim[81], Laptev[89], Meng[105], Wang[181], Farhadi[37], Fathi[39], Holte[64], Jia[72], Jiang[73], Junejo[76], Rodriguez[139], Souvenir[164], Yan[197] | Bag of Keyrames<br>e.g.<br><br>Carlsson[24], Efros[33], Jhuang[71], Thurau[169], Wang[185], Schindler[149], Weinland[186], Zhang[202] |
| Local, Spatial Bag of Features | Feature Grammar<br>e.g.<br><br>Shi[158] | Feature Template<br>e.g.<br><br>Laptev[86], Ke[80] | Bag of ST-Features<br>e.g.<br><br>Schuldt[151], Boiman[16], Dollar[32], Niebles[112], Ikizler[68], Niebles[113], Nowozin[115], Scovanner[152], Wong[193], Filipovych[42], Gilbert[51], Klaser[83], Laptev[87], Liu[96] |

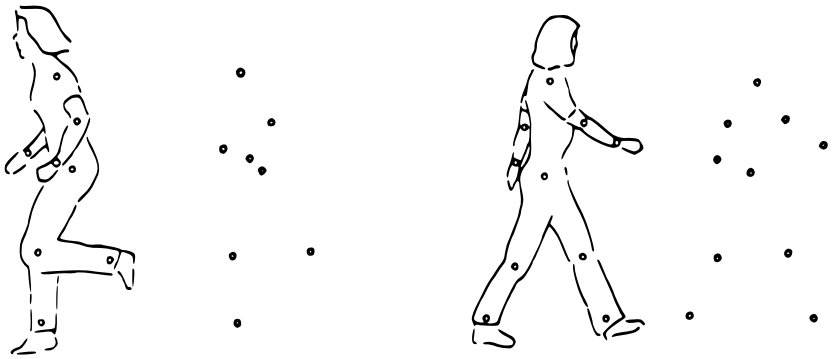


Figure 2: Illustration of moving light displays, taken from [74]. Johansson showed that humans can recognize actions merely from the motion of a few light displays attached to the human body. Awaiting publisher permission

## 2.1 Body models

In this section, we review methods that represent the spatial structure of actions with reference to the human body. In each frame of the observed video stream, the pose of a human body is recovered from a variety of available image features, and action recognition is performed based on such pose estimates. This is an intuitive and biologically-plausible approach to action recognition, which is supported by psychophysical work on visual interpretation of biological motion [74].

Johansson showed that humans can recognize actions merely form the motion of a few moving light displays (MLD) attached to the human body (Figure 2). Over several decades his experiments inspired approaches in action recognition, which used similar representations based on motion of landmark points on the human body. His experiments were also origin of the unresolved controversy on whether humans actually recognize actions directly from 2D motion patterns, or whether they first compute a 3D reconstruction from the motion of the patterns. The observation that upside-down recordings of MLDs are usually not recognized by humans can be interpreted as evidence for the presence of a strong prior model in human perception [166, 52], i.e. humans expect people walking upright and can not easily adapt to strong transformations.

In the context of machine vision, the two approaches have been advocated, resulting in two main classes of methods [107]: 1) *recognition by reconstruction* of 3D body models and 2) *direct recognition* from 2D body models.

**Recognition by reconstruction** divides the task of action recognition in two well separate stages - a motion capture stage which estimate a 3D model of the human body, typically represented as a kinematic joint model; and an action recognition stage which operates on joint trajectories. Two major difficulties are the large number of degrees-of-freedoms of the human body and the high variability of their shapes. As a result, a parametric model of the human body must be carefully selected and calibrated to support action recognition and generalization. A large variety of parametric models have been proposed over the years and we can only mention some of them. See Figure 3 for some examples.

In their early theoretical work on representation of three dimensional shapes [102], Marr and Nishihara proposed a body model consisting of a hierarchy of cylindrical primitives, see Figure 3 a). Such a model was later adopted by several so called *top-down* approaches, e.g. [63, 142]. Top down refers here to approaches that use a 3D model in a generative framework, i.e. 3D body models are generated by sampling from the search space of joint configurations, projected into 2D, and matched against the observation. A more general body model based on super-quadrics was used in the multiview approach [47]. Even more flexible is the model used in [54], which approximates body parts in 3D through a textured spline model. In [133, 67] a *bottom-up* approach is used, which first tracks body parts in 2D, using rectangular appearance patches, and then lifts the tracked 2D configuration into 3D, see Figure 3 c).

Markerless motion capture is a difficult problem in itself, which has been reviewed in other surveys [49, 45, 106, 107, 131] and workshops [61]. MOCAP techniques which require special markers attached to the human have also been used for action recognition. For instance, Campbell et al. compute a joint model from 14 marker points attached to a ballet dancer's body [22], see Figure 3 b). Instead of recovering kinematic joint configurations, several approaches directly work on the trajectories of 3D anatomical landmarks, e.g. head and hand trajectories [21, 19, 192].

**Direct recognition approaches** work from 2D models of the human body, i.e. labeled body parts, without lifting these into 3D. Common 2D representations are stick figures and 2D anatomical landmarks similar to Johansson's MLDs. For instance, Goddard et al. [53] investigate the use of MLDs for action recognition. Guo et al. [58] recover a 2D stick figure from the skeleton of a person's silhouette, see Figure 3 f). Niyogi et al. detect a stick figure from the space-time volume spanned by an image sequence of a walking person [114]. Other direct recognition approaches use coarse 2D body representations based on blobs and patches. For instance [165] detects the hands of a person facing the camera using skin tone based color segmentation, and tracks these over time. Hand and head trajectories have been also used in [19], see Figure 3 d). [195] represents each limb trough a planar patch, and tracks these using optical flow. [20] clusters the human body into a gaussian mixture, based on gradient color and texture values, and models the evolution of the mixture components over time.

To conclude this section, we should note that finding body parts and estimating parametric body models from images remains an unresolved problem, independent of the model used (2D or 3D). Even commercial MOCAP systems using special markers attached to the body rely on heavy user interaction, which makes them unsuitable for recognition tasks, except in very special cases such as automatic annotation of MOCAP data for film productions [6].

Monocular, marker-less MOCAP, which is typically based on difficult non-convex optimizations, is highly prone to such issues as false initialization, convergence to local optima, or non-recovery from failure. Recent methods [143, 161, 1] use strong prior learning to reduce such issues by assuming particular types of activities, walking or running for instance, and thus by imposing strong constraints on the type of possible body configuration. Such prior models hence reduce the search space of possible poses considered, which however limits their application to action recognition [203, 124].

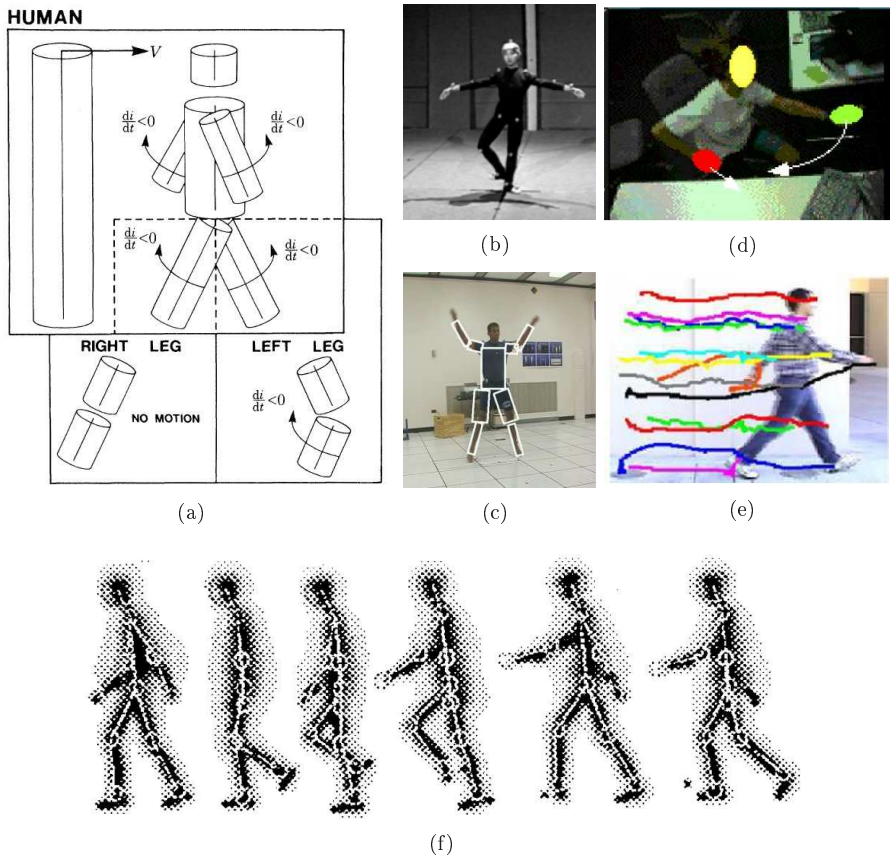



Figure 3: Model based posture representations: (a) hierarchical 3D model based on cylindrical primitives [103]; (b) ballet dancer with markers attached to body [22]; (c) body model based on rectangular patches [133]; (d) blob model [19]; (e) 2D marker trajectories [200]; (f) stick figure [58]. Awaiting publisher permission

## 2.2 Image models

In this section, we review global, image-based representations of actions, also sometimes called *holistic representations*, which do not require the detection and labeling of individual body parts. They only need to detect a region of interest (ROI) centered around the person. In most cases, features are then computed densely on a regular grid bounded by the detected region. As a general term, we call such a representation an *image model* of action. See Figure 4 for some examples.

Image models can be much simpler than parametric body models. As a result, they can be computed more efficiently and robustly. Paradoxically, they have also been shown to be just as discriminative as body models with respect to many classes of actions, when used in combination with temporal representations such as grammars or templates (see following section).

A typical image model is presented by Darell et al. [30], where images of hand gestures are directly correlated, without feature extraction. Their work assumes however a static black background. In most other cases, background subtraction and feature extraction must be performed in a pre-processing stage.

An important class of image models uses silhouettes and contours of the human agent performing the action, As a good example, the seminal works on HMMs for action recognition by Yamato et al. [196], uses silhouette images quantized into super-pixels, each pixel counting the ratio of black and white pixels within its underlying region, as features. A similar representation is also used in [181], see Figure 4 a) and b). In [11] silhouettes are integrated over time in so called *motion history images* (MHI) and *motion energy images* (MEI), see Figure 6 a). In [104] a similar representation is derived based on an *infinite impulse response filter*. [105] proposes the use of a hierarchical MHI. [190] compute 3D *motion history volumes* by lifting silhouettes from multiple view observations into 3D visual hulls, see Figure 6 b). [10] work on the space-time volume spanned by a silhouette sequence over time, see Figure 6 c). Similarly [199] work on a space-time shape, which is computed by tracking a persons contour over time. [116] uses key-frames silhouettes in an HMM and matches them via phase correlation. [137] recognize actions from a condensation filter which uses a spline contours to model the evolution of a person outline over time, see Figure 4 c). [24] compute action characteristic key-frames from shapes derived from canny edge filtered images.

One way to deal with noisy silhouettes, e.g. in outdoor scenes where exact background segmentation is difficult, is to use the chamfer distance [48] as for instance in [35, 186]. Matching of noisy silhouettes can also be made more robust by using phase correlation [116], by stacking the *space-time volume* spanned by silhouette images over time [10, 199], or by using shape context descriptors [100, 162, 202].

As demonstrated by many of the above mentioned approaches, silhouettes provide strong cues for action recognition, and moreover have the advantages of being insensitive to color, texture, and contrast changes. On the downside, silhouette base representations fail in detecting self-occlusions, and depend on a robust background segmentation.

A second important class of image models uses dense optical flow extracted from consecutive images to represent the movements of human agents performing an action. An early example of using optical flow for action recognition is

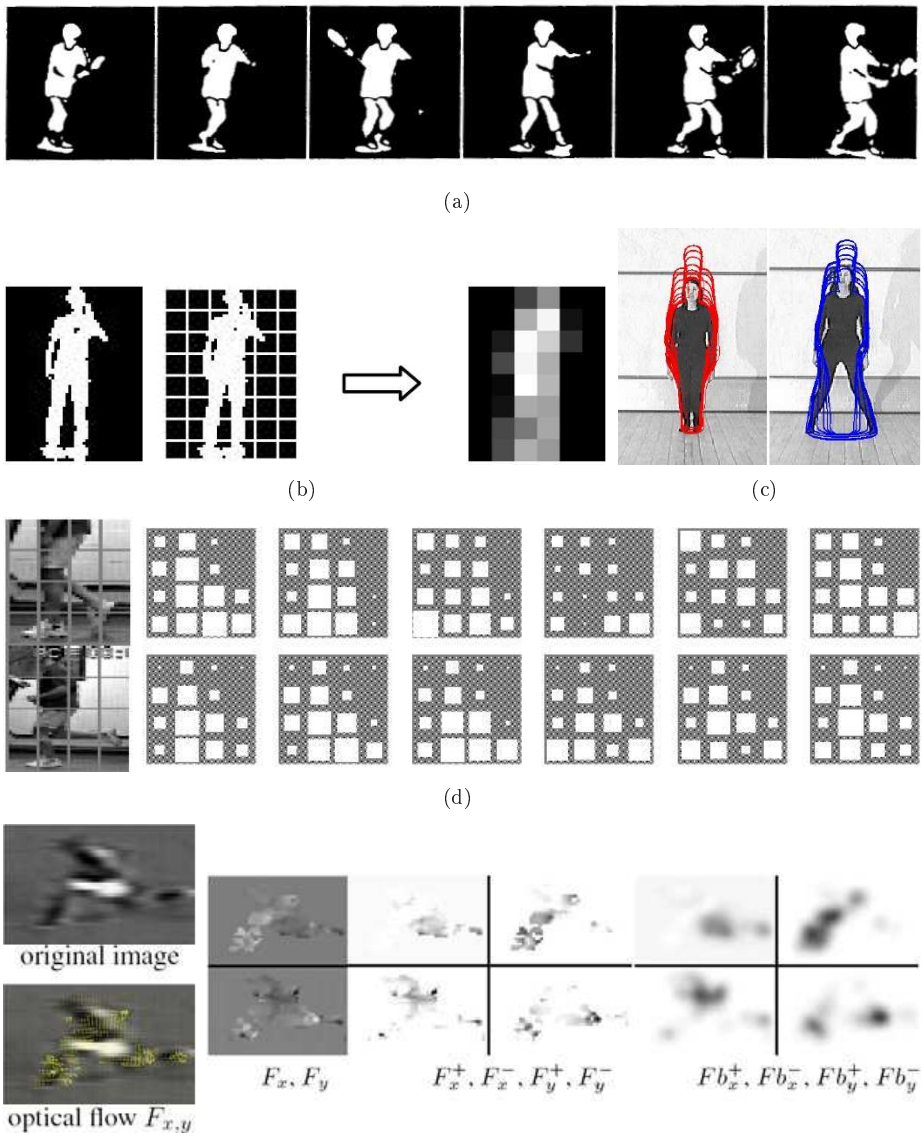



Figure 4: Global posture representations: (a) Silhouettes of tennis strokes [196]; (b) silhouettes pixels accumulated in regular grid [181]; (c) spline contours [137]; (d) optical flow magnitude accumulated in regular grid [130]; (e) optical flow split into directional components, then blurred [33]. Awaiting publisher permission

given by Polana and Nelson [128], where they compute *temporal-textures*, i.e. first and second order statistics based on the direction and magnitude of normal flow, to recognize events such as motion of trees in wind or turbulent motion of water. In [130] Polana and Nelson propose features for human action recognition based on flow magnitudes accumulated in a regular grid of non-overlapping bins, see Figure 4 d). Another early approach which uses optical flow is proposed by Cutler and Turk [27], where the optical flow field is clustered into a set of *motion blobs*, and motion, size, and position of those blobs are used as features for action recognition.

More recently, Efros et al. [33] split the optical flow field into four different scalar fields (corresponding to the negative and positive, horizontal and vertical component of the flow), see Figure 4 e), which are separately matched. This representation was also used in [138, 185].

In the works [79, 89, 39], the adaboost-based Viola–Jones face detector is extended to action recognition by replacing the rectangular image features with statio-temporal cubes computed over optical flow.

Flow based representations do not depend on background subtraction, which makes them more practical than silhouettes in many settings, because they do not require background models. On the downside, they rely on the assumption that image differences can be explained as a result of movement, rather than changes in material properties, lighting, etc.

Another important class of image features is based on gradients. Gradient based features became in particular popular in conjunction with sparse SIFT-like representations (see next section). There are however also several approaches which employ gradients globally. [201] compute gradient fields in $XYT$ direction and represent each frame through the histogram over those gradients. Also the HOG descriptor, which has been very successfully applied to person and object detection [29], has been used for action recognition [170]. Instead of computing a single gradient histogram per frame, the HOG descriptor divides the image grid into regular spaced overlapping blocks, and computes a histogram within each of those blocks.

Gradient features share many properties with optical flow features: they do not depend on background subtraction, but likewise are sensitive to material properties, textures, and lighting, etc. In contrast to optical flow, gradients are discriminative for both moving and non-moving parts, which has advantages as well as disadvantages. For instance static non moving body parts can also provide important cues for an action, but might be easily confused with static object in the background with strong gradients. Recently several works demonstrated superior results by combining gradients and flows [89, 87], or silhouettes and flow [173].

The last class of image models which we discuss, is based on the neuroscientifically inspired HMAX approach [154]. For instance the approaches [71, 149] combine Gabor filters and optical flow in a max-pooling scheme, to simulate the basic stimulus-response functions of a virtual cortex. Because low-frequency Gabor filters can have very similar shapes than oriented gradient filters, both provide similar cues. Yet by using higher-order Gabor filters, additional information can be introduced, which however also leads to a strong increase in computational time.

As explained earlier, image models of actions result in strong simplifications compared to parametric body models. One important consequence is that they

are very sensitive to variations in the view direction of the camera and body sizes of the agent performing the action. It is thus important to account for such variation, either through a large number of different template instances, or by using suitable features and matching functions that are insensitive to such transformations. While holistic approaches have been applied to scenarios of many different kind, they are sometimes advocated as being especially useful for distant-views and coarse representations, e.g. the "30-pixel man" in Efros et al. [33].

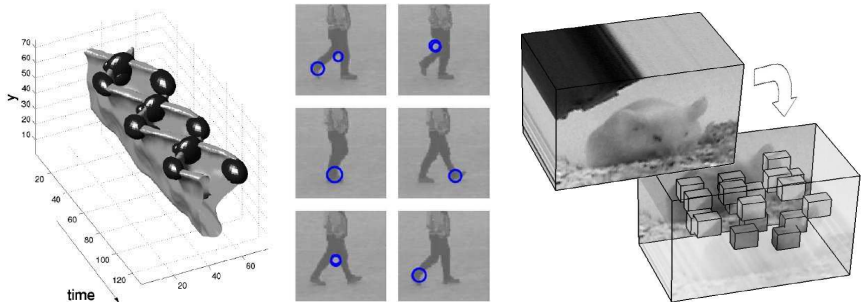
Nonetheless, image-based representations have been used by many approaches of very different kinds. As previously stated, they are more easily extracted than body models while still providing useful cues on actions. Nevertheless, most of the current approaches in this class are based on strong assumptions that need to be addressed in future work. In particular many approaches assume that a ROI around a person, possibly even background subtracted, is provided by a previous processing stage. Consequently, these approaches strongly depend on the progress in related fields such as person detection and tracking. Also, most approaches only operate on fully visible bodies and do not investigate how to adapt global models to partial observations, e.g. occluded bodies or close-up views. Note anyway that video surveillance adapts well to these assumptions since far-views are frequent. Moreover, in such applications additional sensors, including time-of-flight cameras, sonars and tags, can alleviate poor background subtraction or poor motion analysis.

## 2.3 Sparse features

In this section, we review local representations of action which decompose the image/video into smaller interest regions and describe each region as a separate feature. Unlike body model-based representations, the resulting interest regions are, however, *not* linked to certain body parts or even image coordinates. Instead, actions are recognized based on the statistics of those sparse features in the image. An immediate advantage of those approaches is that they neither rely on explicit body part labeling, nor on explicit human detection and localization.

Recently, so called *space-time interest points* [86, 32] became popular, driven by the success of interest points and local descriptors [44, 60, 150, 98] in object recognition and image classification. Such image classification approaches are typically based on bottom up strategies, which first detect interest points in the image, mostly at corner or blob like structures, and then assign each region to a set of preselected vocabulary-features. Image classification reduces then to computations on so called *bag of words*, i.e. histograms that count the occurrence of the vocabulary-features within an image. Similar interest detectors have been proposed by [86] and later by [32], see Figure 5, to locate informative regions in the space-time volumes spanned by video sequences. Typically, for each detected location a compact vector descriptions [86, 32, 152] of the surrounding space-time cuboid is formed and assigned to a set of preselected vocabulary-features. Generally there are many analogies between these approaches and the use of SIFT-descriptors [98] in object recognition.

The work [86] originally extend Harris corner detection [60] and automatic scale selection [93] to 3D space and time. The vocabulary features used in this work are the responses to a set of point-centered and scale-adapted higher order




(a) (b)

Figure 5: Local posture representations: (a) Space-time interest points in [86] are computed at points of high spatiotemporal variation ("spatiotemporal corners"). (b) Spatio-temporal features in [32] are designed to be more responsive then the former space-time interest points. Awaiting publisher permission

gradient filters. This work was extended to BOW and SVM classification in [151] and applied to the meanwhile very well known KTH dataset. [32] proposed an alternative interest point detector based on a quadrature pair of 1D Gabor filters applied temporally and spatially. This work also introduces several SIFT-like [98] space-time descriptors based on local PCA and histogramming of gradient, flow, or brightness values. In [112] an unsupervised approach to learning actions from sparse features is proposed using probabilistic latent semantic analysis [62]. Another interest point detector is proposed in [193], where an image sequence is decomposed into spatial components and motion components using non-negative matrix factorization (NMF). Interest points are then independently detected in 2D spatial and 1D motion space using difference of gaussian (DoG) detectors.

A practical advantage of interest-point approaches is that the detection of the agent need not be performed explicitly for the computation of the space-time features. The detected interest points need to show some consistency for similar observations, but usually they can also account for some outliers. On the downside, the detected features are usually unordered and of variable size, and consequently modeling geometrical and temporal structure is difficult with space-time features. Many approaches stick therefore with the previously mentioned *spatial bags of features* representation, which describes sequences simply through histograms of feature occurrences, hence without modeling any geometrical structure between the feature locations.

Bag-of-feature modeling became prominent in image classification for categorization of objects classes, such as bicycles, cars, and chairs for example. In these settings, discarding global structural information can be even advantageous, as it results in proper insensitivity to interclass variations and view transformations. It is however arguable whether such insensitivity to structural information is advantageous for action recognition, where we are concerned with a single object category, the human body, yet with detailed knowledge about its interclass configuration. Some approaches [113, 194, 42] use graphical models with hidden variables for the position of patches to add structural information to the local features. In [51] so called *compound features* are proposed, which can be seen as some kind of super features taking into account the relative positions

of several features in a neighborhood. Another possibility to add structural information is to divide the image space into several local BOW histograms. For instance the approach [87] computes local features not only at detected interest points, but densely over a space-time volume. BOWs histograms are then computed at multiple positions and scales following a pyramid representation [90]. Though based on local features, such an approach shares as well many of the properties of global image representations such as HOG and HMAX (Section 2.2).

Other interesting issues with BOW based approaches are: how to select a small but discriminative vocabulary [96], and how to combine different types of features, e.g. sparse features and silhouette based features [95].

Finally it is also important to mention, that although most of the previous approaches compute SIFT-like histograms over cubes in 3D space and time, the gradients used in the histograms are nevertheless mostly only 2D spatial. In fact, finding a uniform quantization for vectors on a 3D sphere is a well known problem, which was recently addressed by several papers [152, 83] in the context of deriving 3D SIFT descriptors for action recognition.

Though the majority of sparse feature representations is based on the previously discussed extensions of SIFT, several other sparse/local representations have been proposed. In [80] local patches are computed from an color-based over-segmentation of the space-time volume. Loosely spatial relations between the resulting segments are then learned via pictorial structures [43, 40], and used for matching actions. Also the approach in [16] is based on the idea of learning patches and a loosely spatial configuration of those patches. Even more radical, this approach does not identify patches via segmentation or feature detectors, but searches instead over all possible images patch configurations of a given size.

In summary, sparse features have recently drawn a lot of attention in the action recognition community, probable because they provide many of the advantages that showed successful for static object recognition, and because they are easily and straightforward applied to difficult scenes, e.g. movies or video clips from the internet, that evidently will be very difficult to use with the other kinds of spatial descriptors, i.e. body models and image features. They are less affected by occlusions than global features, and they provide some view invariance to affine transformations. Moreover, they can be used without prior need for person detection and background subtraction. There are however many remaining issues in conjunction with those features, which need to be addressed in future works. For instance it is not clear whether they can be used without person detection in scenes containing multiple persons. Also, they depend on robust detection of the interest points, and consequently seem to work particularly well with fast and periodic motions that provide a lot of reproducible space-time corners, yet it is less obvious how to robustly detect interest points from smooth and monotonic motions. Finally, and as discussed previously, another open issue with those features is how to efficiently add global spatial constraints in to the BOW computation.

# 3 Temporal Action Representations

In the previous section we discussed the different kind of image features that can be extracted from a video sequence to represent the spatial structure of

actions. We will now describe the different representations that can be used to learn the *temporal structure* of actions from such features. As a result, we further classify approaches to action recognition, based on how they express the temporal component of the observations. We distinguish between three main categories of representations: *grammars* are explicit models of the dynamics of actions, which model time as part of the observation, typically through a set of finite states and temporal transitions between these states; *templates* are implicit models of the dynamics of action, where time is frozen into complete temporal blocks of observations; *keyframes* are isolated moments of actions that can be used for recognition without taking into account the temporal relations between them.

## 3.1 Grammars

The approaches discussed in this section represent an action as a sequence of moments, each with their own appearance and dynamics. The choice of dynamic model is generally independent from the type of features used, e.g. a body model (Section 2.1) or image model (Section 2.2). A common way to approximate a dynamical system over feature observations is to group features into similar configurations, i.e. states, and to learn temporal transition functions between these states. Such models fall generally into the class of *state-space* or *graphical models*, which are best described as probabilistic grammars.

Among the versatile probabilistic grammars used for action recognition the most prominent is certainly the *hidden Markov model* (HMM) [132]. The HMM came in particular to fame because of its great success in the speech and natural language processing community. A HMM model of action is typically a probabilistic finite state machine with a single state variable for the moments of the action (e.g. preparation, start, middle and end). State transitions follow the Markovian assumption, i.e. the state at time $t$ only depends on its directly preceding state at time $t-1$. Usually, a separate grammar is learned for each action class, and can use all types of spatial representations as input observations, i.e. body models, image models and sparse features, together or in isolation.

The first work on action recognition using HMMs is probably that of Yamato et al. [196], where a discrete HMM is used to represent sequences over a set of vector quantized silhouette features of tennis footage, see Figure 4 a). Starner et al. [165] use a continuous HMM for recognition of American sign language. Wilson and Bobick [191] are recognizing hand gestures using a HMM. Bregler [20] learns a kind of *switching-state HMM* over a set of autoregressive models, each approximating linear motions of blobs in the video frame.

There are many other approaches using HMMs, to name a few: Brand [17] investigated how a HMM can be learned in one space, e.g. parametric body poses, and mapped to another, e.g. 2D silhouette observation. Wang et al. propose a distance measure between HMMs for unsupervised clustering of gestures. Green and Guan, use HMMs for temporal segmentation of gymnastics footage. Ogale et al. [116] learn HMMs over cluster of key-frame silhouettes, which observe actions from different viewpoints. Lv and Nevatia [99] use HMMs as week classifiers in an adaboost based action recognition approach.

HMMs are purely sequential models of action, which severely limits their use for full-body action recognition, where the different body parts may move independently and in parallel. Various extensions to the more general class

of *dynamic Bayesian networks* (DBN) [50] have been proposed to overcome this limitation. Brand et al. [19] learn coupled HMMs to model interactions between several state variables. They use a two state coupled HMM to recognize interactions between left and right hand motions during Tai Chi exercises. Park et al. [121] use a complex DBN to model interactions between two persons, such as *hugging*, *handshaking*, and *punching* for instance. Peursum et al. [125] model interactions between people and objects in their work using Bayesian networks. N'Guyen et al. [111] propose to use hierarchical HMMs for activity recognition. Lv and Nevatia [100] and Weinland et al. [187] extend HMMs with explicit latent states for view point, to model actions seen from arbitrary views in a single model.

A less obvious limitation of HMMs is that they are *generative models* of actions, which rely on simplifying statistical assumptions for computing the joint probability of the states and the observed features, whereas a more general *discriminative* model may better predict the conditional probability of the states *given* the observed features. As a result, several authors have investigated the use of discriminative models of actions.

Sminchisescu et al. [161] proposes to use Conditional Random Fields (CRF) instead of HMMs. CRFs are discriminative Markov models which can use non-independent features and observations overs time (contrary to the HMM assumption). Furthermore, CRF parameters can be trained to maximize the discriminative power of the classifier, rather than the joint probability of the training examples. Modeling sub-structures within actions is, however, *not* as straightforward with a CRF as with a HMM. The original CRF framework proposed in [161] only modeled dynamics between separate actions instances, but not within single actions. More recent works [182, 108, 181] overcome this issue by using hierarchical layers of latent variables. A CRF approach which explicitly incorporates variables for unknown viewpoint, similar to the HMM approaches [100, 187], is proposed in [110].

Other dynamic models that have been used for action recognition are: auto regressive models [20, 137, 9], context-free grammars [69, 117], general state-space approaches [14], time delayed neural networks [198], feature-structures [84, 169], and Stiefel and Grassmann manifolds[175]. Ali et al. [4] take a more radical path by making the claim that the dynamical processes involved in human action are intrinsically non-linear and non-Markovian. Instead, they propose to model the human body as a chaotic system, to discover the true type of inherent dynamics in a supervised fashion.

A strong advantage of grammars and state-space models is their high degree of modularity. This makes them suitable for generalizing over large variations in acting speeds and styles. Grammars are also compositional, i.e. grammar models of primitive actions can also serve as smaller vocabulary units to build larger networks of complex actions [67], and similarly, complex models can be used to segment sequences into smaller units [18, 54, 123], as discussed more in detail in Section 4.3. Parameters of probabilistic grammars can also be learned quite efficiently, using small numbers of labeled examples in supervised mode, or large numbers of non-labeled examples in non-supervised mode. But the structure of probabilistic grammars must usually be chosen manually (see Kitani et al. [82] for an notable exception). As a result, learning and evaluation of grammar-based action recognition remains an outstanding problem with large numbers of actions classes.

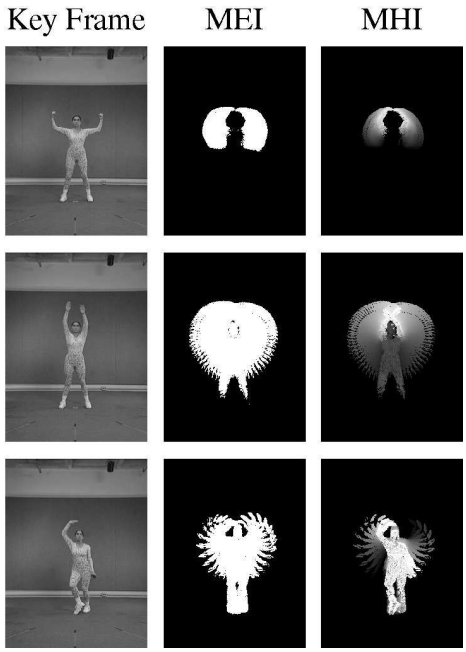
## 3.2 Templates

Instead of representing features and dynamics explicitly and separately in a layered model, some methods encode the dynamics implicitly by directly learning the appearance of complete temporal blocks of features - which we call templates. Typically, template-based approaches directly represent dynamics through example sequences, either by stacking features from several frames into a single feature vector, or by extracting features from the n-dimensional *space-time volume* spanned by a sequence over time, see also Figure 6. Though most of the approaches that use templates are based on image models, such as [10] and [199] that build templates by stacking multiple silhouette images into a single volumetric representation, they can also be used with parametric models, e.g. [58, 114, 195, 136, 55, 5], or even local representations [80].

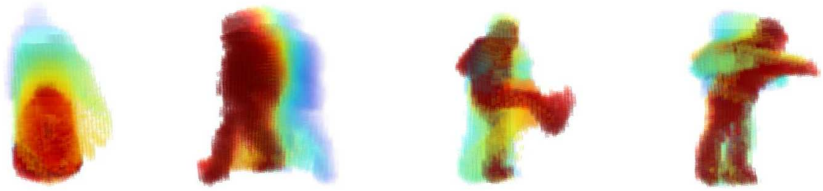
Templates are typically computed over long sequences of frames, and should not be confused with spatio-temporal features or optical flow (Section 2.2 and 2.3), which are computed over small time windows (typically 2-4 frames) and serve as components of other action classifiers, i.e. grammars, templates or keyframes. The seminal work on *action templates* is that of Bobick et al. [11], who build a motion history image (MHI) by mapping successive frames of silhouette sequences into a single image, see Figure 6 a), and extract Hu-moments [65] from this representation. MHIs are generally similar to a depth map computed from a space-time volume. MHIs have been later extended to motion history volumes (MHVs) [188], see Figure 6 b), by using visual hulls computed from multi view sequences instead of the 2D silhouettes used in the original work.

Templates are usually fixed-size vector representations, which makes them straightforward to implement in combination with most static classification techniques. Often simple nearest-neighbor assignment or naive Bayes classification are used in experiments. Contrary to grammars and state-transition models, templates *cannot* represent variations in time, speed, and action style through special variables. Variations are instead implicitly represented through large sets of example sequences, making the classification problem more difficult. In those cases, advanced statical learning methods have been proposed: for instance [58] map joint coordinate sequences onto a low frequency Fourier representation and classify the resulting components using a neural network. [105] proposes to extend MHIs to hierarchical MHIs and classifies those using a support vector machine classifier. The works [89, 79, 39] all share the idea of first computing a flow based spatio-temporal volume, and then using the adaboost classifier to select cuboid based weak classifiers within that volume, similar to the rectangular features of the well known Viola Jones face detector [179].

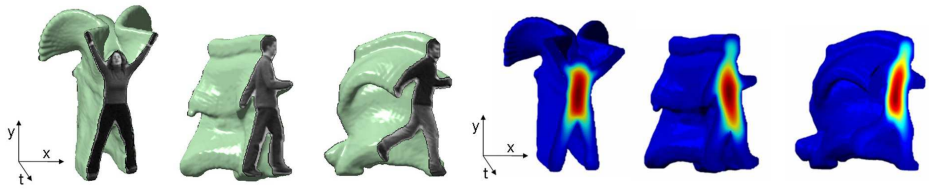
To deal with actions with variable durations, an additional normalization step may be necessary to ensure that the resulting feature vectors have the same dimension, or the more advanced dynamic time warping (DTW) [147] may be used. Darrell et al. [30] correlate observations frames against a set of learned pose templates, and match the resulting sequences of correlation scores using DTW. [114] and [47] use DTW to match sequences of joint model configurations. [178] propose a DTW method for action recognition that allows better modeling of variations within model sequences. Note that DTW is in fact a very simple grammar, where each frame represents a separate state.




(a)



(b)




(c)

Figure 6: Space-time representations: (a) motion energy images (MEI) and motion history images (MHI) [13]; (b) motion history volumes (MHV) [188]; (c) space-time shapes [10]. Awaiting publisher permission

Instead of using multiple templates in a conventional classifier, [139] proposes to build a single template from a collection of templates using a MACH filter, which can then simply be correlated with new observations sequences.

Other important examples of action templates use Fourier or wavelet representations in the temporal domain, [129, 94]. Trajectories of body parts or image features can be also used as templates. For instance, [101] introduce templates of body feature trajectories after tracking over extended time sequences.

In summary, template based representations of very different kind have been proposed. Generally they are effective and discriminative action representations, and in particular attractive because they straightforward integrate with powerful static classifiers such as SVMs or adaboost. Arguably they are more discriminative than Grammars, because they simultaneously constrain space and dynamics, instead of independently treating them on separated layers. Their global characteristic makes them however more difficult adapt to new observations that were not explicitly represented in the training set, e.g. variations in action performance or missing observations because of occlusions. Also in difference to grammars they do not provide internal mechanisms for automatic temporal segmentation (see Section 4.3), and therefore mostly use the more expensive sliding window based search (Section 4.2) for finding good segmentations.

## 3.3 Keyframes

In contrast to previous references, some approaches do not attempt to model the dynamics of action at all. Instead, they attempt to recognize actions from isolated, characteristic *key-frames* or with other time-independent measures such as frequency of feature occurrence over time for instance.

While possibly not practical with all kinds of actions, these approaches have recently made a convincing case that they can be quite effective. An immediate benefit of such representations is a drastic reduction in the complexity and in the sensitivity to temporal variations such as exact length and speed of actions. Another advantage of those methods is that they sometime can even recognize actions from still images, which the human visual system seems to do easily and consistently. Recent methods attempt to emulate this capability by using powerful static matching techniques to compensate for the lack of motion information, which is of course an important cue for action recognition, both by humans and machines, as demonstrated in [74].

Carlsson and Sullivan [24] introduce the use of *key-frames*, i.e. a single characteristic frames of an action, to recognize forehand and backhand strokes in tennis recordings. The term key-frame comes originally from animation and filmmaking, where key frames define start and ending point of a smooth motion. Matching in [24] is based on a sophisticated point to point matching between edge filtered images, to measure the deformation of a edge template with respect to the image observation.

An interesting attempt was made in [149] which does *not* use a single frame, but very short *snippets* of frames, and tries to answer the question of how many frames are required to perform action recognition?

Time independent representations are also important for action/event recognition from single imagery, i.e. photographs. [92] present an approach that combines different visual cues in a generative model to recognize sport events in

static imagery, e.g. *badminton*, *snowboarding*, *sailing*, etc.. Another approach for unsupervised discovery of action classes from single images is proposed in [184].

Besides single static images, sequences can be also encoded without taking temporal relations into account. Histogram techniques, i.e. the so called *bags of words* approaches, have been used to represent sequences simply base on the frequency of feature occurrence e.g. [151, 32, 152, 185]. The biologically motivated system of [71] uses a different technique with feature vectors computed as maximum match responses to a set of prototypes. Similarly, Weinland et al. [186] use an *exemplar-based embedding*, which represents a sequences via its minimum distances to a set of prototypes.

An extension to BOW based on "temporal binning", to explicitly take some short-time temporal ordering of visual words into account is proposed in [115]. Another possibility to add some temporal constraints to BOW based on so-called *spatial-temporal correlograms* was proposed in [148].

Clearly, time independent representations can't be applied to discriminate all kind of actions, e.g. two actions that share similar poses but in different temporal order. They are mostly interesting because they are often simple and efficient to computed, and further time independency results in insensitivity to variations in time scale, e.g. length and speed of an action.

In summary, there is growing interest in using time-independent representations, and in particular the BOW approach. This is partially imposed by the recent trend in using sparse features, which simply seem to work best with such a simplified representation — or at least up to now, no more efficient way of integrating them over time has been found. Nonetheless, and in particular for modeling small atomic motions, dynamic free representations have powerful properties, such as being efficient to compute, insensitive to timescale variation, and nevertheless very discriminative, which makes them as well attractive in a more general context.

# 4 Action Segmentation

In the first two sections of this survey, we have mostly been concerned with approaches that extract visual features from video streams and combine them in space and in time for making a decision on what actions are present in the video. In many cases, those approaches are illustrated with results obtained using *segmented* video clips each showing a single action from start to finish, both for training and testing. But can the two tasks of *action segmentation* and *action recognition* really be performed separatel? There appears to be very little evidence from neuroscience on how motion segmentation and recognition interact in the human visual system. From a computational point of view, it is of course beneficial to segment the video stream before applying recognition, since labeling a segment is much more efficient than labeling all subsegments in a stream. But this raises the difficult issue of finding a generic vocabulary of *parts of actions*, and generic methods for breaking video streams into the corresponding segments. In practice, this appears to be a problem no less difficult than action recognition itself, especially when working with arbitrary views.

We start by discussing different methods used for temporal segmentation. We classify those methods into three broad classes. *Boundary detection* meth-

ods explicitly search for features which characterize start and end points of actions. Such boundaries can be characterized for example as discontinuities or extrema in acceleration, velocities, and curvature. *Sliding window* methods are approaches that evaluate the learned classifiers for all possible segments under a sliding window, and detect segments by searching for peaks in the resulting recognition scores. Compositional approaches are methods which build up representations for sequences of actions, using *higher-level grammars* to model transitions between individual action classes, and dynamic programming techniques to identifying the best possible path through those models. We close this section with a short discussion of methods which attempt to build a vocabulary of generic *action primitives* to simultaneously solve the problems of segmentation and recognition.

## 4.1 Boundary Detection

A common strategy for recognizing actions is to use a generic segmentation method based on detecting motion boundaries, then separately classify the resulting segments. As mentioned earlier, motion boundaries are typically defined as discontinuities and extrema in acceleration, velocities, or curvature of the observed motions. The choice of boundaries thus implicitly results in a basic motion taxonomy.

An early paper by Marr and Vaina [103] discusses the problem of segmenting the 3D movement of the human body, and suggests the use of rest states, i.e. local minima, of the 3D motion of the limbs as natural transitions between primitive movements. Similar, Rubin and Richards define in their work [145] two elementary kinds of motion boundaries: *starts and stops* and *dynamic boundaries*. *Starts and stops* are boundaries that occur whenever a motion changes from a moving state into a rest state and vice versa, they are thus analog to the rest states defined by [103]. Furthermore, *dynamic boundaries* appear between starts and stops and result from discontinuities, e.g. steps or impulses, in force applied to the object in action.

Following this theoretical line of research, computational approaches for motion boundary detection have been proposed [146, 116, 189, 183, 136, 77, 180]. [146] perform an SVD decomposition of a long sequence of optical flow images and detect discontinuities in the trajectories of selected SVD components to segment video into motion patterns. Similar, [116] cluster action sequences by detecting minima and maxima of optical flow inside body silhouettes. Instead of factorizing flow into different components, their method only uses the average global flow magnitude. In [180] impulses in motion, so called *ballistic dynamics*, are modeled using a Bayesian network over various image based features, and used to detect motion boundaries.

In theory, boundary detection methods are attractive because they provide a generic segmentation of the video, which is not dependent on the action classes. In practice, the segmentation must be used with some precautions because (a) they are subject to errors in the recovery of the motion field; (b) they are not stable across view-points; and (c) they are easily confused by the presence of multiple, simultaneous movements.

Extrema in motion magnitude were used in [189] to segment 3D movements of a single actor from visual hulls. That method is independent of viewpoint, and does not require optical flow computations. As a result, it is shown to be

fairly consistent over both training and testing examples, as demonstrated with an unsupervised learning task. Boundary detection methods are most useful in cases where body part trajectories are available. Wang et al. [183] use 2D trajectories of hand gestures and search for local minima of velocities and local maxima of change in direction. Similar [136] examine trajectories of hand motions, see Figure 8(b). [77] use a hierarchical body model and detect local minima in force, momentum, and kinetic energy of the joints. They compare their segmentation results on professional choreographed dances.

## 4.2 Sliding Windows

Another strategy for recognizing actions divides the video sequence into multiple, overlapping segments, using a sliding window. Classification is performed sequentially on all the candidate segments, and peaks in the resulting classification scores are interpreted as action locations. In contrast to boundary detection methods, the segmentation here depends very strongly on the recognition stage. As a result, it should be clear that those methods are not applicable in the training stage. A consistent segmentation of the training examples must be provided manually or through another method, and is a crucial elements for the success or failures of those methods.

A sliding window approach can be used with any of the previously discussed feature representations and classifiers. Many template-based representations (Section 3.2) [201, 205, 41, 79, 80] use a sliding window. Some approaches use them in combination with dynamic time warping (DTW) [30, 109, 5] and even grammars [12, 192].

Compared to boundary detection methods, sliding window methods are usually much more computationally intensive, as they involve many evaluation of all classifiers. To achieve robustness against the duration of actions, they often require multiple window sizes as well, which results in an additional computational burden. Sliding window methods may also produce unpredictable results in the presence of unknown action categories. However, sliding window methods make less assumptions, i.e. they do not assume special boundary criteria, and can be easily integrated on top of any action classifier without requiring further computation of special segmentation features. Indeed, they can be viewed as supervised action segmentation methods, which makes them especially attractive during the learning of novel action classes.

## 4.3 Higher-Level Grammars

In a previous section (Section 3.1), we reviewed representations of individual action classes with grammars, which give a model of the transitions between states (moments) in the action. Similarly, the transitions *between* actions can be modeled with higher-level grammars. This provides a generic method for segmenting sequences of actions. This is a natural strategy when individual actions are already modeled with grammars. Higher-level grammars are built by piecing together the individual models into a single large network, which explicitly models the transition between those single actions. Such networks can be build for instance by joining all models in a common start and end node and by adding a loop-back transition between these two nodes. It is also possible to allow for more complex transitions between actions, e.g. actions may share

states and transitions between actions may be adjusted individually to reflect more realistic the probabilities of one actions following another. Such complex structure are similar to HMM networks used in continuous speech recognition. Segmentation and labeling of a complex action sequence is then computed as a minimum-cost path trough the network using dynamic programming techniques, e.g. the Viterbi path for HMMs [132]. The works [18, 54, 123, 99] use such networks for action recognition based on HMMs. Similar [161, 108] use CRFs, and [158] Semi-Markov models. The work of [137] uses autoregressive models to represent actions, and a condensation filter to switch between these models.

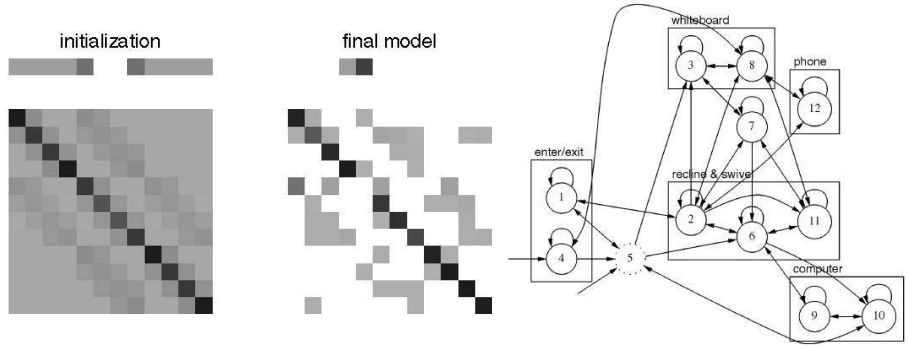
Those approaches make neither the assumptions of the boundary detection methods, nor do they require heavy evaluations such as sliding window approaches. The segmentation is elegantly and efficiently solved using dynamic programming techniques. However, it should be emphasized that learning a higher-level grammar with many actions requires a much larger amount of training data, especially when transitions between actions are learned from real data. In speech recognition such data is available in form of text-documents, word-transcriptions, and phonetically labeled sequences. Similar data does, however, currently not exist for action recognition, and therefore transitions between actions are often set manually, or with strong assumptions, such as uniform transition probabilities.

## 4.4 Action primitives

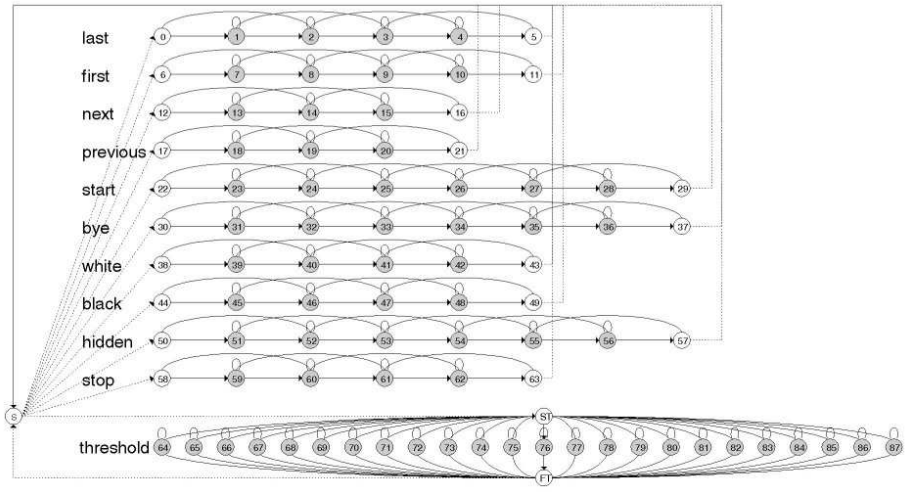
A final class of methods for segmenting actions uses an analogy with speech recognition. All spoken languages are known to be composed of a small number of *phonemes*. Phonemes are the smallest independent units of a language and all utterances in that language can be represented (and recognized) as a composition (sequence) of phonemes. Some authors have made the analogy that at least some actions are *body languages* that can similarly be decomposed into a discrete vocabulary of *primitive actions*. When applicable, this analogy can be used to over-segment the video into a sequence of primitive actions, and leave all the burden of action classification for the next stage of processing.

Independent proposals for motion primitives have been made. [20] introduces *movemes* as the complement to *phonemes* in speech. *Movemes* are basic building blocks of actions words, that can be approximated with a linear system. Similar, [54] chose the name *dyneme*. In their work they build HMM networks based on an empirical chosen alphabets of 35 *dynemes*. Another work which addresses building motor primitives in joint space is [57]. In this work, primitives of kinetic origin are named *kinetemes*. In computer graphics, [144] introduce the concept of *verbs and adverbs* which are used as building blocks to interpolate new motions from example motions. In the work of [6], the user can define a set of motion primitives, which is then used to synthesize new composite motions.

Besides manually defining action taxonomies, several approaches attempt a purely data driven discovery of motion primitives. Brand and Kettnaker [18], see Figure 7(a), start from a fully connected HMM to represent a continuous action sequence. An entropy based minimization is then used to discover independent structures within the HMM by pruning most of the transitions. They evaluate their method on single blob trajectories of office activity and outdoor traffic. [183] segment hand gestures using boundary detection. For each segment a separate HMM is learned and a distance defined between pairs of HMM




(a)




(b)

Figure 7: Primitive discovery and garbage modeling using HMMs: (a) Images from [18]. Transition probabilities and initial state probabilities of a fully connected HMM, (Left) before and (Middle) after structure discovery using entropic estimation. (Right) resulting graphical model (b) HMM network with additional garbage-model (threshold model). The garbage model is build as an ergodic fully connected HMM based on copies of all states in the original models from [91]. Awaiting publisher permission

allows them to hierarchical cluster these HMMs. The work [189] uses boundary detection as well. They use visual hull sequences in 3D and motion energy minima to split those sequences into small action segments. The resulting segments are then agglomerated into a hierarchy of motion clusters. [201] use normalized cuts on correlation matrixes of action sequences to cluster these sequences into different types of motions. [177] propose a SVD techniques for 3rd order tensors, to factorize $\mathbb{R}^{\text{people}\times\text{action}\times\text{time}}$ joint-coordinate tensors into *motion signatures*. [70] use a spatio-temporal extension of the ISOMAP embedding method [168] to discover motor-primitive from MOCAP data.

The composition approach has been extremely successful in the case of facial expressions, where a small number of *Facial Action Units* (FAUs) have become prominent since the work of Ekman and Frisen [34]. It took Ekman decades to build a proper vocabulary and prove that it is indeed generic and stable over different contexts and cultures. Segmentation and recognition of FAUs by computers have been surveyed separately by Fasel and Luettin [38] and can serve as a useful reference for applying composition methods in other areas. Unfortunately, the generalization to *Body Action Units* remain an open problem resisting investigation.

# 5 View-Independent Action Recognition

As mentioned earlier (Section 2.1), fundamental considerations on the model representation, i.e. wether to use a 2D or 3D representation, have a long history in action recognition. Early psychophysical experiments [74] where asking whether humans use structure from motion reconstruction to recognize actions, or whether they recognize actions directly from 2D motion patterns. Approaches demonstrating general qualities of either direction have been proposed.

Following the initial success of those approaches new challenges, such as learning larger number of action classes and robustness under more realistic settings, gained importance. Within this scope a very important demand is independence to viewpoint, which wasn't address by most of the early approaches. It is our opinion, that such considerations bring the issue on how to represent posture, i.e. in 2D or 3D, into a interesting new perspective. In the following discussion we will therefore in particular focus on the different view representations used by action recognition approaches, i.e. 2D, multi-view, or 3D.

We take our taxonomy for view-independent action recognition from work on view-independent shape matching [78], which names three strategies for view-independent matching: normalization, invariance, and exhaustive search. *Normalization* maps observations from different views into a common canonical coordinate frame, matching is then performed under this canonical setting. *Invariance* uses features that do not depend on view transformations, such that the resulting match is the same for any view transformation. *Exhaustive search* takes all possible view transformations into account, and searches for the optimal match within these. All these strategies apply as well to action recognition, with the additional difficulty that the viewpoints themselves may change over time. In the following we discuss approaches based on these strategies, and as mentioned previously, separate as well between view representations in 2D, multi-view, or 3D.

## 5.1 Normalization

In view-normalization based approaches, each observation is mapped to a common canonical coordinate fame. Therefore normalization approaches generally first estimate cues that indicate the transformation from the canonical view frame to the current view of the observation, and then correct the observation with respect to the estimated transformation. Matching then takes place in the normalized coordinate frame.

### 5.1.1 Normalization in 2D

Normalization is used by many approaches as a preprocessing step to remove global scale and translation variations. In particular image models, e.g. silhouette base approaches (Section 2.2), often extract a rectangular region of interest (ROI) around the silhouette, and scale and translate this region to a unit frame. This normalization removes global variations in body size, as well as some scale and translation variations resulting from perspective changes.

Normalization with respect to out-of-plane transformations, e.g. a camera rotation, is not trivial given a single 2D observation. Nevertheless, [140] propose a method, which estimates the 3D orientation of a person from its walking direction in 2D, using knowledge about the ground homography and camera calibration. Assuming only horizontal rotation of the body in 3D, the 2D silhouette of the person is perspectively corrected onto a fronto parallel view and matched against a set of canonical silhouettes.

### 5.1.2 Normalization in 3D

Although it somehow limits the application of action recognition approaches, walking direction as orientation cue was as well used by several 3D based approaches to compute a reference frame for normalization. [15] use multiple views to compute a 3D voxel reconstruction of a walking person. The walking direction is then used to back-project the person silhouette on a view orthogonal to the walking direction, and action recognition is performed on the resulting silhouettes. Also [28] align voxel grids of human bodies using their waking direction. After normalization, they perform action recognition on velocities of body part estimates. [141] extend MHIs [11] to disparity maps. An estimated global flow direction is used in this work to align the 2.5D MHIs. Also several joint model based approaches use an estimated walking direction to estimate the initial model, e.g. [203, 124].

Given a 3D joint body model, an orientation independent joint representation can be compute based on the global body structures. Often the torso is used as reference object to normalize all joints with respect to its orientation. It is further possible to represent each body part with and individual coordinate frames. For example, [47] compute individual reference frames for the torso, arms, and hips.

In summary, normalization approaches are based on the estimation of the body orientation. If strong cues, such as walking direction or a reconstructed body model are available, the orientation can be easily derived. However, all following phases depend on the robustness of this step. Miss alignments, because of noisy estimations or intraclass variations, are likely to affect all following phases of the approach.

## 5.2 View Invariance

View-invariant approaches do not attempt to estimate view transformations between model and observation. Instead view-invariant approaches search for features and matching functions that are independent (i.e. do not change) with respect to the class of view transformations considered.

### 5.2.1 View Invariance in 2D

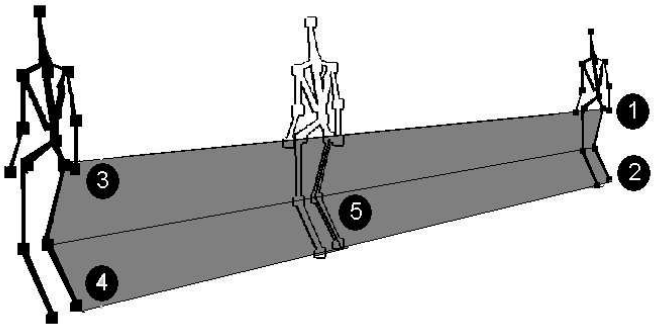
A simple form of view-invariance is based on histogramming. Instead of representing image features in a fixed grid, only the frequency of feature occurrences is stored. Such an representation has been used for instance by [201] to represent distributions of space-time gradients. This representation, however, only provides invariance to translations in the image plane.

The availability of point correspondences, e.g. in form of anatomical landmarks, was frequently used for view-invariant matching between pairs of observations, see Figure 8 for some examples. For instance, an epipolar geometry can be estimated from a subset of point correspondences and then used to constrain the set of all point correspondences, and respectively a matching cost over changing views can be computed without requiring a full 3D reconstruction. I.e. given point matches $(x_i, x'_i)$, $i = 1, \ldots, n \geq 8$ in pairs of images $I, I'$, the fundamental matrix $F$, which holds the relation $x_i F x'_i = 0$, can be estimated. This relation holds however only if all point pairs come from the same rigid object. Hence the resulting residual $\sum_i |x_i F x'_i|^2$ can be used as matching cost [167, 55, 156, 199, 200, 157]. Similar, matrix factorization and rank constraints, as in structure from motion estimation [171], can be used to validate whether point correspondences in two images came from the same single rigid object [153, 136, 134, 135].

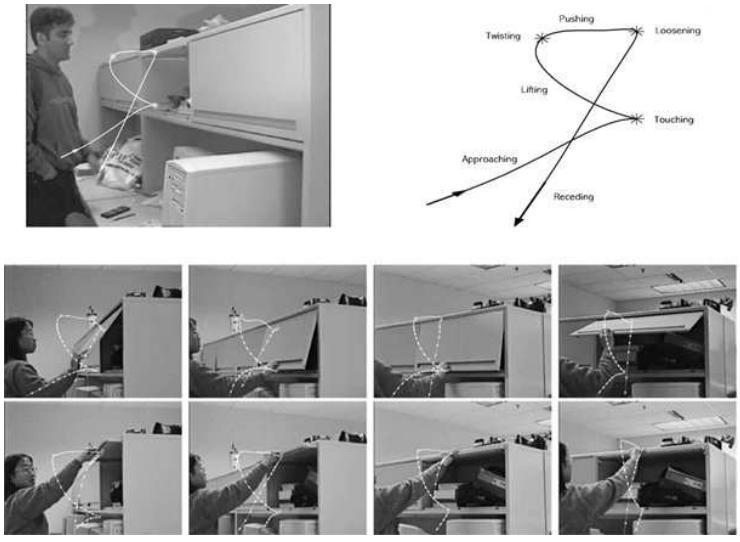
Geometric invariants, i.e. measures that do not change under a geometric transformation, can also been used for invariant matching of landmark points. These invariants can be computed from 5 points that lie in a plane. [118, 119, 120] detect joint configuration during walking cycles that fulfill this condition of 5 landmarks lying in a common plane, and use them to compute geometric invariants.

Note that most of the previously discussed approaches depend on detection of point correspondences, which is very difficult in practice. Further note, that although these approaches are single view 2D, computing the fundamental matrix and structure from motion factorizations are already first steps towards a 3D reconstruction.

More recently, an invariant approach that optional uses point correspondences or image features without correspondences is proposed in [76], where actions are learned from arbitrary number of views by extracting view-invariant features based on frame-to-frame self-similarities within a sequence. The representation discards all information related to an absolute reference frame, it does not depend on position or appearance, and only is based on the relative change between frames. It is shown in [76], that such features remain surprisingly stable under changing viewing conditions. The approach [37] is purely based on image observations. It provides some kind of view-invariance by using a *transfer learning* approach, which maps an action model from a *source-view* into a novel *target-views*. To establish such a transfer mapping, explicit samples of corresponding observations from source and target view must be available during learning, those need however not necessarily provide views of the same action class, for which the transfer function is learned.




(a)




(b)

Figure 8: View invariant action recognition: (a) geometrical invariants can be computed from 5 point that lie in a plane [118]; (b) View-invariant matching of hand trajectories [135]. Point matches between different observations are computed from discontinuities in motion trajectories. Awaiting publisher permission

### 5.2.2 View Invariance in 3D

Based on 3D body part trajectories [21] investigates 10 different view-invariant representations in their work. These include shift invariant velocities $(dx, dy, dz)$ in cartesian coordinates, and shift and horizontal rotation invariant velocities $(dr, d\theta, dz)$ in polar coordinate. In evaluation on 18 Tai Chi gestures, the polar coordinate representation has best overall recognition rates.

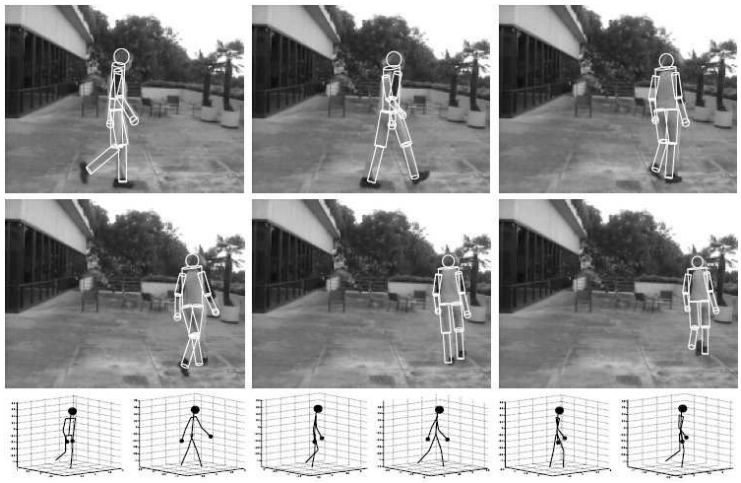
There are few other view-invariant approaches in 3D, and especially few approaches that do not depend on a joint body model or point correspondences. An exception is the work of [26], which proposes a view invariant pose representation based on a voxel reconstruction. Cylindrical 3D histograms similar to the 2D shape context descriptor [7], are used as invariant measure of the voxel's distribution in this work. The same descriptor was later used by [126] for action recognition. Another invariant representation based on 3D shape-context and spherical harmonics was proposed in [64]. [190] propose a view-invariant representation in 3D, based on Fourier coefficients in cylindrical coordinates, applies to a 3D extension of MHIs [13]. The same representation was later used in [175], but with a more sophisticated modeling approach based on Stiefel and Grassmann manifolds. Also the paper [23] proposes a view-invariant representation based on MHIs in 3D. They use a invariant representation based on 3D moments [97].

## 5.3 Exhaustive Search

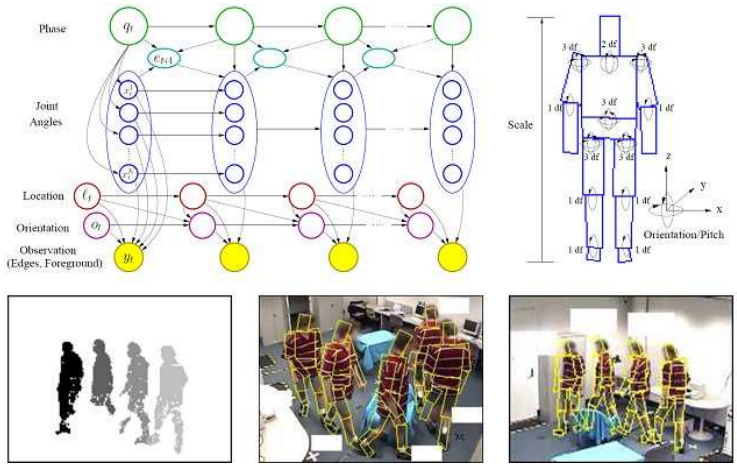
Instead of deciding on a single transformation, as it is typical for normalization methods, or discarding all transformation dependent information, as with invariant methods, one can search over all possible transformations considered. At first sight, an exhaustive search may seem heavy on computational resources. Yet, reasonable assumptions, such as restrictions to certain classes of transformations, advanced search strategies, and propagation of findings over time, can drastically reduce the search space. Moreover, with the steadily increasing performance of modern computer systems, the computational expense of such methods is about to become fairly manageable.

### 5.3.1 Exhaustive Search using Multiple 2D Views

Several approaches use a fixed set of cameras installed around the actor, and simultaneously record the actions from this multiple views. During recognition, an observation is then matched against each recorded view and the best matching pair is identified. In their work on MHIs, [13] record actions with 7 cameras, each with an offset of 30° in the horizontal plane around the actors. During recognition two cameras with 90° offset are used, and matched against all pairs of recorded views with the same 90° offset. An action is then labeled with respect to the best average match of two cameras. Similar [116, 117] use 8 prerecorded views, and a single view during recognition. Their work uses a single HMM to mode temporal relations between prototype silhouettes and view changes over time. Consequently, they can recognize smooth view changes, e.g. a person slowly turning around while performing an action. The work [3] use as well 8 prerecorded views and HMMs. The individual view HMMs are however learned without transitions between close views.



(a)




(b)

Figure 9: Generative MOCAP: (a) Tracking a person walking in cycle [159]. (b) Factor-state hierarchical HMM body model and tracking results using the generative approach in [124]. Awaiting publisher permission

### 5.3.2 Exhaustive Search using a 3D Model

To achieve more flexibility with respect to changes in camera setup, an internal model based on a 3D representation can be used. From such a 3D representation, and given camera parameters, any possible 2D view observation can be rendered. Such *generative approaches* are frequently used in MOCAP, where parameterized 3D models of the human body are projected into 2D. These models have explicit variables for global 3D position and orientation, that are estimated simultaneously with the remaining joint parameters, e.g. [31, 159, 124], see also Figure 9.

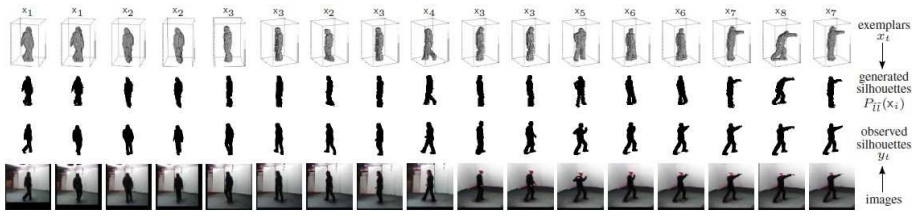



Figure 10: Generative model based on set of exemplary 3D key-poses. Used for view-independent action recognition in [187]. Awaiting publisher permission

Similar methods haven been proposed for action recognition, and extended such that they do not require a joint model. The approach [100] uses a small set of synthetic 3D key-poses, rendered from a modeling software. Actions are then matched against the poses, which are projected into 2D with respect to all possible view transformations. Dynamics over poses and changes in view transformations are modeled in a dynamic network, and the best pose-view sequence is found via a dynamic programming search. The approach [187], see Figure 10, shares as well the idea of projecting a set of learned 3D key-poses into 2D to infer actions from arbitrary view. This work uses a HMMs with additional capabilities to model unknown view transformations, a transformed HMM [46, 75, 172]. Such a HMM allows to compute view transformation independent observation probabilities by marginalizing over all possible transformations. Computing marginals is indeed a form of exhaustive search, except that no deterministic decision are made. Instead a probability taking all possible search results into account is computed. Another approach based on the idea of projecting a 3D model into 2D to achieve view-independence is proposed in [110]. This approach extends previously discussed works by using a conditional random fields (CRF) instead of the HMM, and by extending feature observations to include not only silhouettes, but as well optical flow.

Instead of projecting a 3D model into 2D, the opposite direction is taken in [197], where features are detected first in 2D, and then back-projected onto *4D action shapes*. This approach requires as well an optimization over the possible view orientations to find the best 2D-3D matching.

Another direction is takin in [164]. Instead of using pose exemplars and explicitly producing the 3D to 2D projection, they directly learn a function, which takes as input the viewpoint and outputs the corresponding silhouette representation. They use therefore a silhouette representation based on the Radon-transform and cubic B-splines to represent those silhouettes as a function of viewpoint. Given a query action, they search for the viewpoint that minimizes the residual between the learned function and the silhouette representation of the query action.

# 6 Datasets

Finally, we want to discuss some of the datasets, which are currently used by many of the action recognition approaches as a benchmark. Unfortunately acquiring realistic action footage, including ground truth data, is a very difficult

Figure 11: Example images from the KTH dataset

and time consuming task, and currently there is certainly lack of such data in action recognition.

The three popular datasets, which are currently used by most of the approaches are: *KTH* [151], *Weizmann* [10], and *IXMAS* [190]. They all contain around 6-11 action performed by various actors. They are all not very realistic and share strong simplifying assumptions, such as static background, no occlusions, given temporal segmentation, and only a single actor.

The recognition rates of the papers discussed in this survey on those datasets are given in Table 2. It is however important to note that not all approaches follow the exactly same evaluation methodologies, so approaches can't be compared purely based on those results. Moreover, in light of the simplifying assumptions made in the datasets, it is not evident how those results might extrapolated to more complex scenarios. The three datasets are detailed in the following.

## 6.1 The KTH Dataset

The KTH dataset [151], Figure 11, contains the six actions *walking, jogging, running, boxing, hand waving and hand clapping*, performed several times by 25 subjects in four different scenarios. Overall it contains 2391 sequences. It has fewer action classes than the two other datasets, but the most samples per class. It is hence well suited for learning intensive approaches, e.g. approaches based on SVMs.

In difference to the two other datasets it does not provide background models and extracted silhouettes, and moreover some of the scenes are recorded with a slightly shaking camera. Most approaches that evaluate on the KTH dataset are hence based on sparse features (Section 2.3) which are best suited to such scenarios. Recently, however some approaches which require background subtraction, e.g. [71, 149], reported as well results on the dataset, and the authors [71] published some low-level foreground masks for the dataset on their webpage. Not surprisingly, that those works are among the best performing approaches on the dataset. The original paper [151] reported a recognition rate of 71.7% on the dataset. More recently several approaches reported recognition rates above 90% up to 97%.

## 6.2 The Weizmann dataset

The Weizmann dataset [10], Figure 12, contains the nine actions *running, walking, bending, jumping-jack, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, galloping-sideways, waving-two-hands, waving-one-hand*, performed by nine different actors. The dataset's website also provides a tenth action *skip*,

Figure 12: Example images from the Weizmann dataset

which is however not used by all approaches. Overall it contains 93 sequences, all performed in front of similar plain backgrounds, and with a static camera. It is the smallest of the three datasets considered.

The original approach [10] reported already a very high recognition rate of 99.6%, and similar results have been archived by many of the subsequent approaches. It appears hence to be the easiest of the three datasets, and result do not provide much insight into the quality of an approach. Nevertheless it is still used in recent works.

Generally approaches which use the background subtracted silhouettes achieve best rates (up to 100%). Recently also several approaches that only depend on person location, but not on extracted silhouettes could report very high recognition rates, e.g. [149, 186]. The lowest rates are archived with sparse feature based approaches, apparently because they neither use person location nor background subtraction.

## 6.3 The IXMAS dataset

The INRIA XMAS dataset [190], Figure 13, contains the 11 daily-life actions: *check watch*, *cross arms*, *scratch head*, *sit down*, *get up*, *turn around*, *walk*, *wave*, *punch*, *kick*, *pick-up*, performed each 3 times by 11 non-professional actors. Note that there are two more actors and actions on the dataset's website, but those have not been used by most of the approaches. The actions were filmed with 5 carefully calibrated and synchronized cameras. Overall it contains hence 429 multi-view sequences, or, if the views are considered individually, 2145 sequences. It also provides background subtracted silhouettes and reconstructed visual hulls.

The scenes are recorded in front of simple static studio-like backgrounds. Its main difficulty comes from the changing viewpoint, that is caused by the different camera configurations and the fact that actors freely chose their orientation while performing the actions. Respectively, the dataset is in particular used by view-independent approaches (Section 5).

The best known recognition rates were recently reported by [175] (98.78%) using 3D MHVs [190] and a modeling approach based on Stiefel and Grassmann Manifolds. Approaches which use only a single camera for recognition reported results up to 82%.

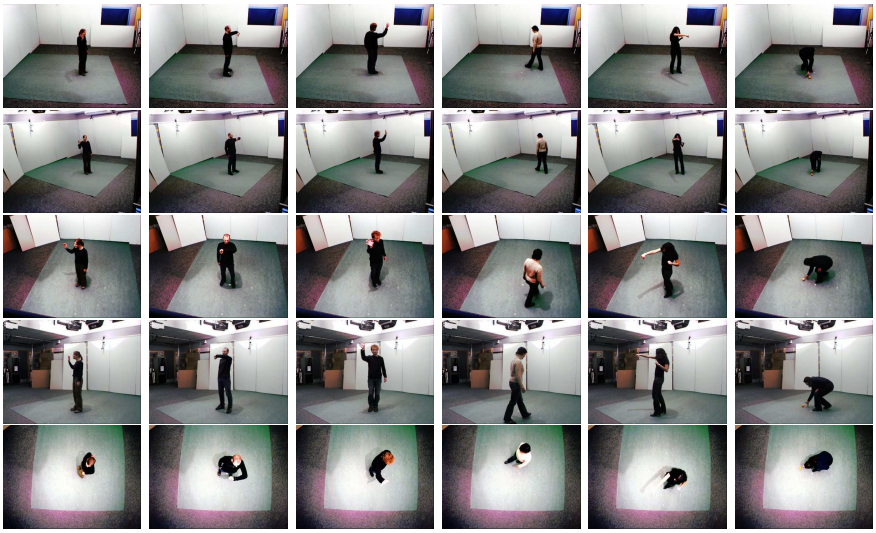


Figure 13: Example images from the five view used in the IXMAS dataset

## 6.4 Other datasets

There are several other not so frequently used datasets available. The CMU MoBo database [56] and the HUMAN-EVA database [160] were primarily designed for motion capture and hence only contain very simple action. Nevertheless they have been used by some approaches in action recognition. The MuHAVI dataset from Kingston University in London contains 17 action classes performed by 14 non-professional actors. They were filmed by 8 surveillance cameras. The cameras are not synchronized or calibrated. From the same research group, the ViHaSi [59] dataset contains synthetic sequences of silhouettes generated from MOCAP data in Autodesk MotionBuilder. The HOHA datasets [88] are a large collection of short segments of real *Hollywood* movies, annotated with 12 action classes: *answer-phone, drive-car, eat, fight-person, get-out-car, hand-shake, hug-person, kiss, run, sit-down, sit-up, stand-up*. The actions are performed by professional actors, under a wide range of camera viewpoints and in very different styles. This is a very challenging dataset, including inter-actions with people (fight-person, hand-shake, hug-person, kiss) and objects (answer-phone, drive-car, getout-car), which are outside the scope of this survey.

Table 2: We list the papers discussed in this survey with respect to the spatial representation used (spatial): body model (bm), image model (im), or sparse features (sf); with respect to the temporal model (temp) is used: grammar (gr), templates (tmp), key-frames (key), or bag of words (bow); with respect to the temporal segmentation (temp) that is used: boundary detection (bd), sliding window (sw), grammar (gr), or action primitives (ap); with respect to the invariant representation that is used: normalization (norm), invariance (inv), or exhaustive search (exh); and finally we also show the best average recognition rate that a paper reported for the three benchmark datasets: KTH, Weizmann (Weiz), and IXMAS (IXS). A $*$ means thereby that an approach was evaluated on a certain dataset, but did not report results in terms of average recognition rate.

| year | paper | spatial | temp | segm | view | KTH | Weiz | IXS |
|---|---|---|---|---|---|---|---|---|
| 1978 | Marr [102] | bm | - | - | - | - | - | - |
| 1982 | Marr [103] | bm | - | bd | - | - | - | - |
| 1985 | Rubin [145] | - | - | bd | - | - | - | - |
| 1989 | Goddard [52] | bm | gr | bd | - | - | - | - |
| 1992 | Polana [128] | im | tmp | - | - | - | - | - |
| 1992 | Yamato [196] | im | gr | - | - | - | - | - |
| 1993 | Darrell [30] | im | tmp | - | - | - | - | - |
| 1994 | Guo [58] | bm | tmp | - | - | - | - | - |
| 1994 | Niyogi [114] | bm | tmp | - | - | - | - | - |
| 1994 | Polana [130] | im | tmp | pd | - | - | - | - |
| 1995 | Campbell [22] | bm | gr | sw | - | - | - | - |
| 1995 | Gavrila [47] | bm | tmp | - | norm | - | - | - |
| 1996 | Bobick [11] | im | tmp | sw | - | - | - | - |
| 1996 | Campbell [21] | bm | gr | - | inv | - | - | - |
| 1997 | Bobick [14] | bm | gr | sw | - | - | - | - |
| 1997 | Brand [19] | bm | gr | - | - | - | - | - |
| 1997 | Bregler [20] | bm | gr | sw | - | - | - | - |
| 1997 | Seitz [153] | bm | tmp | - | inv | - | - | - |
| 1998 | Bobick [12] | bm | gr | sw | - | - | - | - |
| 1998 | Yacoob [195] | bm | tmp | - | - | - | - | - |
| 1999 | Brand [17] | im | gr | gr | exh | - | - | - |
| 1999 | Rittscher [137] | im | gr | gr | - | - | - | - |
| 2000 | Brand [18] | bm | gr | gr | - | - | - | - |
| 2000 | Rui [146] | - | - | bd | - | - | - | - |
| 2001 | Bissacco [9] | bm | gr | - | - | - | - | - |
| 2001 | Bobick [13] | im | tmp | sw | exh | - | - | - |
| 2001 | Carlsson [24] | im | key | sw | - | - | - | - |
| 2001 | Syeda-Mahmood [167] | im | tmp | - | inv | - | - | - |
| 2001 | Wang [183] | bm | gr | bd | - | - | - | - |
| 2001 | Zelnik-Manor [201] | im | tmp | sw | - | - | - | - |
| 2002 | Ben-Arie [8] | bm | tmp | - | - | - | - | - |
| 2002 | Jenkins [70] | bm | - | ap | - | - | - | - |
| 2002 | Kojima [84] | bm | gr | - | - | - | - | - |
| 2002 | Rao [136] | bm | tmp | bd | inv | - | - | - |
| 2002 | Vasilescu [177] | bm | - | ap | - | - | - | - |
| 2002 | Zhao [203] | bm | gr | - | norm | - | - | - |
| 2003 | Bodor [15] | im | tmp | - | norm | - | - | - |
| 2003 | Cohen [26] | im | - | - | inv | - | - | - |
| 2003 | Efros [33] | im | key | sw | - | - | - | - |
| 2003 | Elgammal [35] | im | gr | - | - | - | - | - |
| 2003 | Kahol [77] | bm | - | bd | - | - | - | - |
| 2003 | Laptev [86] | sf | tmp | - | - | - | - | - |
| 2003 | Masoud [104] | im | tmp | - | - | - | - | - |
| 2003 | Parameswaran [118] | bm | tmp | - | inv | - | - | - |
| 2003 | Park [121] | bm | gr | - | - | - | - | - |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2003 | Ramanan [133] | bm | gr | - | exh | - | - | - |
| 2003 | Rao [135] | bm | tmp | - | inv | - | - | - |
| 2003 | Rao [134] | bm | tmp | - | inv | - | - | - |
| 2003 | Cuzzolin [28] | im | gr | - | norm | - | - | - |
| 2004 | Green [54] | bm | gr | ap | - | - | - | - |
| 2004 | Gritai [55] | bm | tmp | - | inv | - | - | - |
| 2004 | Ogale [116] | im | gr | - | exh | - | - | - |
| 2004 | Schuldt [151] | sf | bow | - | - | 71.7 | - | - |
| 2004 | Zhong [205] | im | tmp | sw | - | - | - | - |
| 2005 | Alon [5] | bm | tmp | sw | - | - | - | - |
| 2005 | Blank [10] | im | tmp | sw | - | - | 99.6 | - |
| 2005 | Boiman [16] | sf | key | sw | - | - | - | - |
| 2005 | Dollar [32] | sf | bow | - | - | 81.2 | - | - |
| 2005 | Feng [41] | im | tmp | sw | - | - | - | - |
| 2005 | Ke [79] | im | tmp | sw | - | 63.0 | - | - |
| 2005 | Nguyen [111] | bm | gr | gr | - | - | - | - |
| 2005 | Ogale [117] | im | gr | - | exh | - | - | - |
| 2005 | Peursum [125] | bm | gr | sw | - | - | - | - |
| 2005 | Robertson [138] | im | gr | sw | - | - | - | - |
| 2005 | Shechtman [155] | im | tmp | sw | - | - | - | - |
| 2005 | Sheikh [156] | bm | tmp | - | inv | - | - | - |
| 2005 | Sminchisescu [161] | im | gr | gr | - | - | - | - |
| 2005 | Smith [163] | im | tmp | - | - | - | - | - |
| 2005 | Weinland [188] | im | tmp | - | inv | - | - | - |
| 2005 | Yilmaz [199] | im | tmp | - | inv | - | - | - |
| 2005 | Yilmaz [200] | bm | tmp | - | inv | - | - | - |
| 2006 | Ahmad [3] | im | gr | - | exh | - | - | - |
| 2006 | Canton-Ferrer [23] | im | tmp | bd | inv | - | - | - |
| 2006 | Kitani [82] | bm | gr | - | - | - | - | - |
| 2006 | Lv [99] | bm | gr/tmp | gr | - | - | - | - |
| 2006 | Niebles [112] | sf | bow | - | - | 81.5 | - | - |
| 2006 | Pierobon [126] | im | tmp | - | inv | - | - | - |
| 2006 | Rogez [140] | im | - | - | norm | - | - | - |
| 2006 | Roh [141] | im | tmp | - | norm | - | - | - |
| 2006 | Veeraraghavan [178] | im | tmp | - | - | - | - | - |
| 2006 | Wang [184] | bm | gr | sw | - | - | - | - |
| 2006 | Weinland [190] | im | tmp | bd | inv | - | - | 93.3 |
| 2006 | Weinland [189] | im | tmp | bd | inv | - | - | 93.3 |
| 2007 | Ali [4] | bm | gr | - | - | - | 92.6 | - |
| 2007 | Guerra-Filho [57] | bm | gr | pa | exh | - | - | - |
| 2007 | Ikizler [68] | sf | bow | - | - | - | 100 | - |
| 2007 | Ikizler [67] | bm | gr | gr | exh | - | - | - |
| 2007 | Jhuang [71] | im | key | - | - | 96.0 | 98.8 | - |
| 2007 | Ke [80] | sf | tmp | sw | - | - | - | - |
| 2007 | Kim [81] | im | tmp | - | - | 95.3 | - | - |
| 2007 | Laptev [89] | im | tmp | sw | - | - | - | - |
| 2007 | Li [92] | sf | - | - | - | - | - | - |
| 2007 | Lv [100] | im | gr | - | exh | - | - | 80.6 |
| 2007 | Meng [105] | im | tmp | - | - | 80.3 | - | - |
| 2007 | Morency [108] | bm | gr | gr | - | - | - | - |
| 2007 | Niebles [113] | sf | bow | - | - | - | 72.8 | - |
| 2007 | Nowozin [115] | sf | bow | - | - | 84.7 | - | - |
| 2007 | Peursum [124] | bm | gr | - | norm | - | - | - |
| 2007 | Scovanner [152] | sf | bow | - | - | - | 82.6 | - |
| 2007 | Thurau [169] | im | bow | - | - | - | 86.7 | - |
| 2007 | Turaga [176] | im | gr | bd | - | - | - | - |
| 2007 | Wang [181] | im | tmp | - | - | - | 100 | - |
| 2007 | Wang [185] | im | bow | - | - | 92.4 | - | - |
| 2007 | Weinland [187] | im | gr | - | exh | - | - | 81.3 |
| 2007 | Wong [193] | sf | bow | - | - | 81.0 | - | - |
| 2008 | Farhadi [37] | im | tmp | - | inv | - | - | 58.1 |
| 2008 | Fathi [39] | im | tmp | - | - | 90.5 | 100 | - |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2008 | Filipovych [42] | sf | bow | - | - | - | 88.9 | - |
| 2008 | Gilbert [51] | sf | bow | - | - | 89.9 | - | - |
| 2008 | Holte [64] | im | tmp | - | inv | - | - | - |
| 2008 | Jia [72] | im | tmp | - | - | - | 90.9 | - |
| 2008 | Jiang [73] | im | tmp | - | - | - | * | - |
| 2008 | Junejo [76] | im/bm | tmp | - | inv | - | 95.3 | 72.7 |
| 2008 | Klaser [83] | sf | bow | - | - | 91.4 | 84.3 | - |
| 2008 | Laptev [87] | sf | bow | - | - | 91.8 | - | - |
| 2008 | Liu [96] | sf | bow | - | - | 94.2 | - | 82.8 |
| 2008 | Liu [95] | im/sf | bow | - | - | - | 89.3 | 78.5 |
| 2008 | Natarajan [110] | im | gr | - | exh | - | - | - |
| 2008 | Rodriguez [139] | im | tmp | - | - | - | * | - |
| 2008 | Schindler [149] | im | key | - | - | 96.7 | 100 | - |
| 2008 | Shen [157] | bm | tmp | - | inv | - | - | - |
| 2008 | Shi [158] | sf | gr | gr | - | - | - | - |
| 2008 | Souvenir [164] | im | tmp | - | exh | - | - | * |
| 2008 | Turaga [175] | im | gr | - | - | - | - | 98.8 |
| 2008 | Thurau [170] | im | bow | - | - | * | 94.4 | - |
| 2008 | Tran [173] | im | key | - | - | - | 100 | 81 |
| 2008 | Vitaladevuni [180] | im | gr | bd | - | - | - | 87.0 |
| 2008 | Weinland [186] | im | bow | - | - | - | 100 | - |
| 2008 | Yan [197] | im | tmp | - | exh | - | - | 78.0 |
| 2008 | Zhang [202] | im | bow | - | - | 81.3 | 92.9 | - |

# 7 Conclusion

In this paper we have given a survey of work in action recognition. We have classified approaches with respect to how they represent the spatial and temporal structure of actions, how they segment actions from an input stream, and how they provide a view-invariant representation of actions. We could identify a large body of different proposals and interesting contributions — overall we could compile approx. 200 papers for this survey. There are still some important issues, which we did not find addressed in most of those works, such as for instance: scalability of action recognition systems, handling of unknown motions, dealing with occlusions, and scenes containing multiple persons. Those are very important issues, not only in action recognition but for most cutting edge research topics within computer vision, e.g. face or object recognition, and it will remain very challenging for the community to come up with solutions for each of them.

A problem that we could in particular identify for action recognition is the lack of available realistic datasets. Obviously, acquiring realistic annotated video footage is a very time consuming and complex tasks — for action recognition probably even more than for related disciplines, e.g. object or face recognition. As a consequence, and as also discussed in the previous section, the current benchmark datasets in action recognition are limited to up to 11 different actions, which are performed by specially instructed actors and recorded in scenarios with very strong simplifying assumptions, such as plain background, fixed camera setups, and given temporal segmentation. It is thus not any longer surprising when recent works report results close to perfect on those recordings, but very little can be gained from over-simplified situations, e.g. where actions are limited to walking and running motions. Working on true surveillance footage, sport recordings, movies, and video data from the internet, will help us to discover the real requirements for action recognition, and it will help us to shift focus to the other important issues involved in action recognition, such as

previously discussed segmentation of continuous actions, dealing with unknown motions, multiple persons, and view invariance.

# References

[1] A. Agarwal, B. Triggs, Recovering 3d human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 44–58.

[2] J. K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding 73 (3) (1999) 428–440.

[3] M. Ahmad, S.-W. Lee, Hmm-based human action recognition using multiview image sequences, in: International Conference on Pattern Recognition, vol. 1, 2006, pp. 263–266.

[4] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, in: IEEE International Conference on Computer Vision, 2007.

[5] J. Alon, V. Athitsos, S. Sclaroff, Accurate and efficient gesture spotting via pruning and subgesture reasoning, in: International Workshop on Human-Computer Interaction, 2005, pp. 189–198.

[6] O. Arikan, D. A. Forsyth, J. F. O'Brien, Motion synthesis from annotations, ACM Transactions on Graphics 22 (3) (2003) 402–408.

[7] S. Belongie, J. Malik, Matching with shape contexts, in: IEEE Workshop on Content-based Access of Image and Video Libraries, 2000, pp. 20–26.

[8] J. Ben-Arie, Z. Wang, P. Pandit, S. Rajaram, Human activity recognition using multidimensional indexing, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (8) (2002) 1091–1104.

[9] A. Bissacco, A. Chiuso, Y. Ma, S. Soatto, Recognition of human gaits, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2001, pp. II–52–II–57 vol.2.

[10] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: IEEE International Conference on Computer Vision, 2005, pp. 1395–1402.

[11] A. Bobick, J. Davis, Real-time recognition of activity using temporal templates, in: Workshop on Applications of Computer Vision, 1996, pp. 39–42.

[12] A. Bobick, Y. Ivanov, Action recognition using probabilistic parsing, in: IEEE Conference on Computer Vision and Pattern Recognition, 1998, pp. 196–202.

[13] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.

[14] A. F. Bobick, A. D. Wilson, A state-based approach to the representation and recognition of gesture, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (12) (1997) 1325–1337.

[15] R. Bodor, B. Jackson, O. Masoud, N. Papanikolopoulos, Image-based reconstruction for view-independent human motion recognition, in: International Conference on Intelligent Robots and Systems, vol. 2, 2003, pp. 1548–1553 vol.2.

[16] O. Boiman, M. Irani, Detecting irregularities in images and in video, in: IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 462–469 Vol. 1.

[17] M. Brand, Shadow puppetry, in: IEEE International Conference on Computer Vision, 1999, pp. 1237–1244.

[18] M. Brand, V. Kettnaker, Discovery and segmentation of activities in video, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 844–851.

[19] M. Brand, N. Oliver, A. Pentland, Coupled hidden markov models for complex action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 994–999.

[20] C. Bregler, Learning and recognizing human dynamics in video sequences, in: IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 568–574.

[21] L. W. Campbell, D. A. Becker, A. Azarbayejani, A. F. Bobick, A. Pentland, Invariant features for 3-d gesture recognition, in: IEEE International Conference on Automatic Face and Gesture Recognition, 1996, pp. 157–163.

[22] L. W. Campbell, A. F. Bobick, Recognition of human body motion using phase space constraints, in: IEEE International Conference on Computer Vision, 1995, pp. 624–630.

[23] C. Canton-Ferrer, J. R. Casas, M. Pardàs, Human model and motion based 3d action recognition in multiple view scenarios (invited paper), in: European Signal Processing Conference, 2006.

[24] S. Carlsson, J. Sullivan, Action recognition by shape matching to key frames, in: Workshop on Models versus Exemplars in Computer Vision, 2001.

[25] C. Cedras, M. Shah, Motion-based recognition: A survey, Image and Vision Computing 13 (2) (1995) 129–155.

[26] I. Cohen, H. Li, Inference of human postures by classification of 3d human body shape, in: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003, pp. 74–81.

[27] R. Cutler, M. Turk, View-based interpretation of real-time optical flow for gesture recognition, in: IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 416–421.

[28] F. Cuzzolin, A. Sarti, S. Tubaro, Action modeling with volumetric data, in: International Conference on Image Processing, vol. 2, 2004, pp. 881–884 Vol.2.

[29] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 886–893 vol. 1.

[30] T. Darrell, A. Pentland, Space-time gestures, in: IEEE Conference on Computer Vision and Pattern Recognition, 1993, pp. 335–340.

[31] J. Deutscher, A. Blake, I. Reid, Articulated body motion capture by annealed particle filtering, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2000, pp. 126–133 vol.2.

[32] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: International Workshop on Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.

[33] A. A. Efros, A. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: IEEE International Conference on Computer Vision, 2003, pp. 726–733.

[34] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978.

[35] A. M. Elgammal, V. D. Shet, Y. Yacoob, L. S. Davis, Learning dynamics for exemplar-based gesture recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 571–578.

[36] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, X. Twombly, Vision-based hand pose estimation: A review, Computer Vision and Image Understanding 108 (1-2) (2007) 52–73.

[37] A. Farhadi, M. K. Tabrizi, Learning to recognize activities from the wrong view point, in: European Conference on Computer Vision, 2008, pp. 154–166.

[38] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition 36 (1) (2003) 259 – 275.

[39] A. Fathi, G. Mori, Action recognition by learning mid-level motion features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[40] P. Felzenszwalb, D. Huttenlocher, Efficient matching of pictorial structures, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2000, pp. 66–73 vol.2.

[41] Z. Feng, T.-J. Cham, Video-based human action classi.cation with ambiguous correspondences, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, p. 82.

[42] R. Filipovych, E. Ribeiro, Learning human motion models from unsegmented videos, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.

[43] M. Fischler, R. Elschlager, The representation and matching of pictorial structures, IEEE Transactions on Computers 22 (1) (1973) 67–92.

[44] W. Forstner, E. Gulch, A fast operator for detection and precise location of distinct points, corners and centres of circular features, in: Intercommission Conference on Fast Processing of Photogrammetric Data, 1987, pp. 281–305.

[45] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, D. Ramanan, Computational studies of human motion: part 1, tracking and motion synthesis, Found. Trends. Comput. Graph. Vis. 1 (2-3) (2005) 77–254.

[46] B. J. Frey, N. Jojic, Learning graphical models of images, videos and their spatial transformations, in: Conference on Uncertainty in Artificial Intelligence, 2000, pp. 184–191.

[47] D. Gavrila, L. Davis, Towards 3-d model-based tracking and recognition of human movement, in: International Workshop on Face and Gesture Recognition, 1995, pp. 272–277.

[48] D. Gavrila, V. Philomin, Real-time object detection for smart vehicles, in: IEEE International Conference on Computer Vision, 1999, pp. 87–93.

[49] D. M. Gavrila, The visual analysis of human movement: A survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98.

[50] Z. Ghahramani, Learning dynamic Bayesian networks, Lecture Notes in Computer Science 1387 (1998) 168–197.

[51] A. Gilbert, J. Illingworth, R. Bowden, Scale invariant action recognition using compound features mined from dense spatio-temporal corners, in: European Conference on Computer Vision, 2008, pp. I: 222–233.

[52] N. H. Goddard, The interpretation of visual motion: recognizing moving light displays, in: Workshop on Visual Motion, 1989, pp. 212 – 220.

[53] N. H. Goddard, The perception of articulated motion: Recognizing moving light displays, Ph.D. thesis, University of Rochester, Rochester, NY, USA (1992).

[54] R. D. Green, L. Guan, Quantifying and recognizing human movement patterns from monocular video images-part i: a new framework for modeling human motion., IEEE Transactions on Circuits and Systems for Video Technology 14 (2) (2004) 179–190.

[55] A. Gritai, Y. Sheikh, M. Shah, On the use of anthropometry in the invariant analysis of human actions, in: International Conference on Pattern Recognition, vol. 2, 2004, pp. 923–926 Vol.2.

[56] R. Gross, J. Shi, The cmu motion of body (mobo) database, Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA (June 2001).

[57] G. Guerra-Filho, Y. Aloimonos, A language for human action, Computer 40 (5) (2007) 42–51.

[58] Y. Guo, G. Xu, S. Tsuji, Understanding human motion patterns, in: International Conference on Pattern Recognition, vol. 2, 1994, pp. 325–329.

[59] P. R. H. Ragheb, S. Velastin, T. Ellis, Vihasi: Virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods, in: Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, 2008.

[60] C. Harris, M. Stephens, A combined corner and edge detector, in: Alvey Conference, 1988, pp. 147–152.

[61] A. Hilton, P. Fua, R. Ronfard, Modeling people: vision-based understanding of a person's shape, appearance, movement, and behaviour, Computer Vision and Image Understanding 104 (2) (2006) 87–89.

[62] T. Hofmann, Probabilistic latent semantic analysis, in: Uncertainty in Artificial Intelligence, 1999, pp. 289–296.

[63] D. Hogg, Model-based vision: A program to see a walking person, Image and Vision Computing 1 (1) (1983) 5–20.

[64] M. B. Holte, T. B. Moeslund, P. Fihl, View invariant gesture recognition using the csem swissranger camera, Int. J. Intell. Syst. Technol. Appl. 5 (3/4) (2008) 295–303.

[65] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory 8 (1962) 179–187.

[66] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Transactions on Systems, Man and Cybernetics 34 (2004) 334–352.

[67] N. Ikizler, , D. Forsyth, Searching video for complex activities with finite state models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[68] N. Ikizler, P. Duygulu, Human action recognition using distribution of oriented rectangular patches, in: Workshop on Human Motion Understanding, Modeling, Capture and Animation, 2007.

[69] Y. A. Ivanov, A. F. Bobick, Recognition of visual activities and interactions by stochastic parsing, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 852–872.

[70] O. Jenkins, M. Mataric, Deriving action and behavior primitives from human motion data, in: International Conference on Intelligent Robots and System, vol. 3, 2002, pp. 2551–2556 vol.3.

[71] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action, in: IEEE International Conference on Computer Vision, 2007.

[72] K. Jia, D.-Y. Yeung, Human action recognition using local spatio-temporal discriminant embedding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[73] H. Jiang, D. R. Martin, Finding actions using shape flows, in: European Conference on Computer Vision, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 278–292.

[74] G. Johansson, Visual perception of biological motion and a model for its analysis, Perception & Psychophysics 1414 (2) (1973) 201–211.

[75] N. Jojic, N. Petrovic, B. Frey, T. Huang, Transformed hidden markov models: Estimating mixture models of images and inferring spatial transformations in video sequences, in: IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 26–33.

[76] I. Junejo, E. Dexter, I. Laptev, P. Pérez, Cross-view action recognition from temporal self-similarities, in: European Conference on Computer Vision, Marseille, France, 2008.

[77] K. Kahol, P. Tripathi, S. Panchanathan, T. Rikakis, Gesture segmentation in complex motion sequences, in: International Conference on Image Processing, vol. 2, 2003, pp. II–105–8 vol.3.

[78] M. Kazhdan, Shape representations and algorithms for 3d model retrieval, Ph.D. thesis, Princeton University (April 2004).

[79] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 166–173.

[80] Y. Ke, R. Sukthankar, M. Hebert, Event detection in crowded videos, in: IEEE International Conference on Computer Vision, 2007.

[81] T. Kim, S. Wong, R. Cipolla, Tensor canonical correlation analysis for action classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[82] K. M. Kitani, Y. Sato, A. Sugimoto, An mdl approach to learning activity grammars, in: Korea-Japan Joint Workshop on Pattern Recognition, 2006.

[83] A. Kläser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference, 2008.

[84] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions., International Journal of Computer Vision 50 (2) (2002) 171–184.

[85] V. Krüger, D. Kragic, A. Ude, C. Geib, The meaning of action: a review on action recognition and mapping, Advanced Robotics 21 (13) (2007) 1473–1501.

[86] I. Laptev, T. Lindeberg, Space-time interest points, in: IEEE International Conference on Computer Vision, 2003, pp. 432–439 vol.1.

[87] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[88] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision & Pattern Recognition, 2008.

[89] I. Laptev, P. Pérez, Retrieving actions in movies, in: IEEE International Conference on Computer Vision, 2007.

[90] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.

[91] H.-K. Lee, J. Kim, An hmm-based threshold model approach for gesture recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (10) (1999) 961–973.

[92] L.-J. Li, L. Fei-Fei, What, where and who? classifying events by scene and object recognition, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[93] T. Lindeberg, On automatic selection of temporal scales in time-causal scale-space, in: International Workshop on Algebraic Frames for the Perception-Action Cycle, London, UK, 1997, pp. 94–113.

[94] F. Liu, R. Picard, Finding periodicity in space and time, in: IEEE International Conference on Computer Vision, 1998, pp. 376–383.

[95] J. Liu, S. Ali, M. Shah, Recognizing human actions using multiple features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[96] J. Liu, M. Shah, Learning human actions via information maximization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[97] C. Lo, H. Don, 3-d moment forms: Their construction and application to object identification and positioning, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (10) (1989) 1053–1064.

[98] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[99] F. Lv, R. Nevatia, Recognition and segmentation of 3-d human action using hmm and multi-class adaboost, in: European Conference on Computer Vision, 2006, pp. 359–372.

[100] F. Lv, R. Nevatia, Single view human action recognition using key pose matching and viterbi path searching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[101] F. Lv, R. Nevatia, M. Lee, 3d human action recognition using spatio-temporal motion templates, in: ICCV Workshop on Human-Computer Interaction, 2005, p. 120.

[102] D. Marr, H. K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes, Philosophical Transactions of the Royal Society of London B 200 (1140) (1978) 269–294.

[103] D. Marr, L. Vaina, Representation and recognition of the movements of shapes, Philosophical Transactions of the Royal Society of London B 214 (1982) 501–524.

[104] O. Masoud, N. Papanikolopoulos, A method for human action recognition, Image and Vision Computing 21 (8) (2003) 729–743.

[105] H. Meng, N. Pears, C. Bailey, A human action recognition system for embedded computer vision application, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.

[106] T. B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding 81 (3) (2001) 231–268.

[107] T. B. Moeslund, A. Hilton, V. Krûger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2) (2006) 90–126.

[108] L.-P. Morency, A. Quattoni, T. Darrell, Latent-dynamic discriminative models for continuous gesture recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[109] P. Morguet, M. Lang, Spotting dynamic hand gestures in video image sequences using hidden markov models, in: International Conference on Image Processing, 1998, pp. 193–197 vol.3.

[110] P. Natarajan, R. Nevatia, View and scale invariant action recognition using multiview shape-flow models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[111] N. Nguyen, D. Phung, S. Venkatesh, H. Bui, Learning and detecting activities from movement trajectories using the hierarchical hidden markov model, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 955–960 vol. 2.

[112] J. Niebles, H. Wang, H. Wang, L. Fei Fei, Unsupervised learning of human action categories using spatial-temporal words, in: British Machine Vision Conference, 2006, p. III:1249.

[113] J. C. Niebles, L. Fei-Fei, A hierarchical model of shape and appearance for human action classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[114] S. Niyogi, E. Adelson, Analyzing and recognizing walking figures in xyt, in: IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 469–474.

[115] S. Nowozin, G. Bakir, K. Tsuda, Discriminative subsequence mining for action classification, in: IEEE International Conference on Computer Vision, 2007.

[116] A. Ogale, A. Karapurkar, G. Guerra-Filho, Y. Aloimonos, View-invariant identification of pose sequences for action recognition., in: VACE, 2004.

[117] A. S. Ogale, A. Karapurkar, Y. Aloimonos, View-invariant modeling and recognition of human actions using grammars, in: Workshop on Dynamical Vision, 2005, pp. 115–126.

[118] V. Parameswaran, R. Chellappa, View invariants for human action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2003, pp. II–613–19 vol.2.

[119] V. Parameswaran, R. Chellappa, Human action-recognition using mutual invariants, Computer Vision and Image Understanding 98 (2) (2005) 295–325.

[120] V. Parameswaran, R. Chellappa, View invariance for human action recognition, International Journal of Computer Vision 66 (1) (2006) 83–101.

[121] S. Park, J. K. Aggarwal, Recognition of two-person interactions using a hierarchical bayesian network, in: ACM SIGMM International Workshop on Video Surveillance, 2003, pp. 65–76.

[122] V. I. Pavlovic, R. Sharma, T. S. Huang, Visual interpretation of hand gestures for human-computer interaction: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 677–695.

[123] P. Peursum, H. Bui, S. Venkatesh, G. West, Human action segmentation via controlled use of missing data in hmms, in: International Conference on Pattern Recognition, vol. 4, 2004, pp. 440–445 Vol.4.

[124] P. Peursum, S. Venkatesh, G. West, Tracking-as-recognition for articulated full-body human motion analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[125] P. Peursum, G. West, S. Venkatesh, Combining image regions and human activity for indirect object recognition in indoor wide-angle views, in: IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 82–89 Vol. 1.

[126] M. Pierobon, M. Marcon, A. Sarti, S. Tubaro, 3-d body posture tracking for human action template matching, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 2006, pp. II–II.

[127] C. Pinhanez, Representation and recognition of action in interactive spaces, Ph.D. thesis, MIT Media Lab (1999).

[128] R. Polana, R. Nelson, Recognition of motion from temporal texture, in: IEEE Conference on Computer Vision and Pattern Recognition, 1992, pp. 129–134.

[129] R. Polana, R. Nelson, Detecting activities, in: IEEE Conference on Computer Vision and Pattern Recognition, 1993, pp. 2–7.

[130] R. Polana, R. Nelson, Low level recognition of human motion (or how to get your man without finding his body parts), in: NAM, 1994.

[131] R. Poppe, Vision-based human motion analysis: An overview, Computer Vision and Image Understanding 108 (1-2) (2007) 4–18, iSSN=1077-3142.

[132] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (1990) 267–296.

[133] D. Ramanan, D. A. Forsyth, Automatic annotation of everyday movements, Tech. Rep. UCB/CSD-03-1262, EECS Department, University of California, Berkeley (Jul 2003).

[134] C. Rao, A. Gritai, M. Shah, T. Syeda-Mahmood, View-invariant alignment and matching of video sequences, in: IEEE International Conference on Computer Vision, 2003, pp. 939–945 vol.2.

[135] C. Rao, M. Shah, T. Syeda-Mahmood, Invariance in motion analysis of videos, in: ACM International conference on Multimedia, ACM, New York, NY, USA, 2003, pp. 518–527.

[136] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, International Journal of Computer Vision 50 (2) (2002) 203–226.

[137] J. Rittscher, A. Blake, Classification of human body motion, in: IEEE International Conference on Computer Vision, 1999, pp. 634–639.

[138] N. Robertson, I. Reid, Behaviour understanding in video: A combined method, in: IEEE International Conference on Computer Vision, 2005, pp. 808–815.

[139] M. D. Rodriguez, J. Ahmed, M. Shah, Action mach a spatio-temporal maximum average correlation height filter for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[140] G. Rogez, J. Guerrero, J. Martinez del Rincon, C. Orrite Urunuela, Viewpoint independent human motion analysis in man-made environments, in: British Machine Vision Conference, 2006, p. II:659.

[141] M.-C. Roh, H.-K. Shin, S.-W. Lee, S.-W. Lee, Volume motion template for view-invariant gesture recognition, in: International Conference on Pattern Recognition, vol. 2, 2006, pp. 1229–1232.

[142] K. Rohr, Towards model-based recognition of human movements in image sequences, Graphical Model and Image Processing 59 (1) (1994) 94–115.

[143] R. Rosales, S. Sclaroff, Inferring body pose without tracking body parts, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2000, pp. 721–727 vol.2.

[144] C. Rose, M. Cohen, B. Bodenheimer, Verbs and adverbs: multidimensional motion interpolation, IEEE Computer Graphics and Applications 18 (5) (1998) 32–40.

[145] J. M. Rubin, W. A. Richards, Boundaries of visual motion, Tech. rep., Massachusetts Institute of Technology, Cambridge, MA, USA (1985).

[146] Y. Rui, P. Anandan, Segmenting visual actions based on spatio-temporal motion patterns, in: IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 1111–1118.

[147] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 26 (1) (1978) 43–49.

[148] S. Savarese, A. DelPozo, J. C. Niebles, L. Fei-Fei, Spatial-temporal correlatons for unsupervised action classification, in: IEEE Workshop on Motion and video Computing, 2008, pp. 1–8.

[149] K. Schindler, L. van Gool, Action snippets: How many frames does human action recognition require?, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[150] C. Schmid, R. Mohr, C. Bauckhage, Evaluation of interest point detectors, International Journal of Computer Vision 37 (2) (2000) 151–172.

[151] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local svm approach, in: International Conference on Pattern Recognition, 2004, pp. 32–36.

[152] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: ACM International conference on Multimedia, 2007, pp. 357–360.

[153] S. M. Seitz, C. R. Dyer, View-invariant analysis of cyclic motion, International Journal of Computer Vision 25 (3) (1997) 231–251.

[154] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 994–1000 vol. 2.

[155] E. Shechtman, E. Shechtman, M. Irani, Space-time behavior based correlation, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 405–412 vol. 1.

[156] M. Sheikh, M. Shah, Exploring the space of a human action, in: IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 144–149.

[157] Y. Shen, H. Foroosh, View-invariant action recognition using fundamental ratios, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–6.

[158] Q. Shi, L. Wang, L. Cheng, A. Smola, Discriminative human action segmentation and recognition using semi-markov model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[159] H. Sidenbladh, M. J. Black, D. J. Fleet, Stochastic tracking of 3d human figures using 2d image motion, in: European Conference on Computer Vision, Springer-Verlag, London, UK, 2000, pp. 702–718.

[160] L. Sigal, M. J. Black, Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Tech. rep., Brown University (2006).

[161] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Conditional models for contextual human motion recognition, in: IEEE International Conference on Computer Vision, vol. 2, 2005, pp. 1808–1815 Vol. 2.

[162] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Discriminative density propagation for 3d human motion estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 390–397.

[163] P. Smith, N. da Vitoria Lobo, M. Shah, Temporalboost for event recognition, in: IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 733–740 Vol. 1.

[164] R. Souvenir, J. Babbs, Learning the viewpoint manifold for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.

[165] T. Starner, A. Pentland, Real-time american sign language recognition from video using hidden markov models, in: International Symposium on Computer Vision, 1995, pp. 265–270.

[166] S. Sumi, Upside-down presentation of the johansson moving light-spot pattern, Perception 13 (3) (1984) 283–286.

[167] T. Syeda-Mahmood, M. Vasilescu, S. Sethi, Recognizing action events from multiple viewpoints, in: EventVideo01, 2001, pp. 64–72.

[168] J. B. Tenenbaum, V. Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

[169] C. Thurau, Behavior histograms for action recognition and human detection, in: Workshop on Human Motion Understanding, Modeling, Capture and Animation, 2007, pp. 299–312.

[170] C. Thurau, V. Hlavac, Pose primitive based human action recognition in videos or still images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[171] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, International Journal of Computer Vision 9 (2) (1992) 137–154.

[172] K. Toyama, A. Blake, Probabilistic tracking in a metric space., in: IEEE International Conference on Computer Vision, 2001, pp. 50–59.

[173] D. Tran, A. Sorokin, Human activity recognition with metric learning, in: European Conference on Computer Vision, 2008.

[174] P. Turaga, R. Chellappa, V. S. Subrahmanian, O. Udrea, Machine recognition of human activities: A survey, IEEE Transactions on Circuits and Systems for Video Technology 18 (11) (2008) 1473–1488.

[175] P. Turaga, A. Veeraraghavan, R. Chellappa, Statistical analysis on stiefel and grassmann manifolds with applications in computer vision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[176] P. K. Turaga, A. Veeraraghavan, R. Chellappa, From videos to verbs: Mining videos for activities using a cascade of dynamical systems, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[177] M. Vasilescu, Human motion signatures: analysis, synthesis, recognition, in: International Conference on Pattern Recognition, vol. 3, 2002, pp. 456–460 vol.3.

[178] A. Veeraraghavan, R. Chellappa, A. Roy-Chowdhury, The function space of an activity, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 959–968.

[179] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. 511–518.

[180] S. Vitaladevuni, V. Kellokumpu, L. Davis, Action recognition using ballistic dynamics., in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, p. 8 p.

[181] L. Wang, D. Suter, Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[182] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, T. Darrell, Hidden conditional random fields for gesture recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1521–1527.

[183] T.-S. Wang, H.-Y. Shum, Y.-Q. Xu, N.-N. Zheng, Unsupervised analysis of human gestures, in: IEEE Pacific Rim Conference on Multimedia, Springer-Verlag, London, UK, 2001, pp. 174–181.

[184] Y. Wang, H. Jiang, M. Drew, Z.-N. Li, G. Mori, Unsupervised discovery of action classes, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1654–1661.

[185] Y. Wang, P. Sabzmeydani, G. Mori, Semi-latent dirichlet allocation: A hierarchical model for human action recognition, in: Workshop on Human Motion Understanding, Modeling, Capture and Animation, 2007.

[186] D. Weinland, E. Boyer, Action recognition using exemplar-based embedding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[187] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, in: IEEE International Conference on Computer Vision, 2007.

[188] D. Weinland, R. Ronfard, E. Boyer, Motion history volumes for free viewpoint action recognition, in: IEEE International Workshop on modeling People and Human Interaction, 2005.

[189] D. Weinland, R. Ronfard, E. Boyer, Automatic discovery of action taxonomies from multiple views, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[190] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding 104 (2-3) (2006) 249–257.

[191] A. Wilson, A. Bobick, Learning visual behavior for gesture analysis, in: International Symposium on Computer Vision, 1995, pp. 229–234.

[192] A. D. Wilson, A. F. Bobick, Parametric hidden markov models for gesture recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (9) (1999) 884–900.

[193] S.-F. Wong, R. Cipolla, Extracting spatiotemporal interest points using global information, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[194] S.-F. Wong, T.-K. Kim, R. Cipolla, Learning motion categories using both semantic and structural information, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.

[195] Y. Yacoob, M. Black, Parameterized modeling and recognition of activities, in: IEEE International Conference on Computer Vision, 1998, pp. 120–127.

[196] J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden markov model, in: IEEE Conference on Computer Vision and Pattern Recognition, 1992, pp. 379–385.

[197] P. Yan, S. M. Khan, M. Shah, Learning 4d action feature models for arbitrary view action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[198] M.-H. Yang, N. Ahuja, Recognizing hand gesture using motion trajectories, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 1999, pp. –472 Vol. 1.

[199] A. Yilmaz, M. Shah, Actions sketch: A novel action representation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. I: 984–989.

[200] A. Yilmaz, M. Shah, Recognizing human actions in videos acquired by uncalibrated moving cameras, in: IEEE International Conference on Computer Vision, 2005, pp. 150–157.

[201] L. Zelnik-Manor, M. Irani, Event-based video analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[202] Z. Zhang, Y. Hu, S. Chan, L.-T. Chia, Motion context: A new representation for human action recognition, in: European Conference on Computer Vision, 2008, pp. 817–829.

[203] T. Zhao, R. Nevatia, 3d tracking of human locomotion: a tracking as recognition approach, in: International Conference on Pattern Recognition, vol. 1, 2002, pp. 546–551.

[204] W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfeld, Face recognition: A literature survey, ACM Computing Surveys 35 (4) (2003) 399–458.

[205] H. Zhong, J. Shi, M. Visontai, Detecting unusual activity in video, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2004, pp. II–819–II–826 Vol.2.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399